

# Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization

Gen Li      Wei Fan      Yuting Wei

Department of Statistics and Data Science, the Wharton School  
University of Pennsylvania, Philadelphia, PA

February 7, 2023

## Abstract

This paper is concerned with the problem of reconstructing an unknown rank-one matrix with prior structural information from noisy observations. While computing the Bayes-optimal estimator seems intractable in general due to its nonconvex nature, Approximate Message Passing (AMP) emerges as an efficient first-order method to approximate the Bayes-optimal estimator. However, the theoretical underpinnings of AMP remain largely unavailable when it starts from random initialization, a scheme of critical practical utility. Focusing on a prototypical model called  $\mathbb{Z}_2$  synchronization, we characterize the finite-sample dynamics of AMP from random initialization, uncovering its rapid global convergence. Our theory provides the first non-asymptotic characterization of AMP in this model without requiring either an informative initialization (e.g., spectral initialization) or sample splitting.

**Keywords:** approximate message passing, random initialization, non-asymptotic analysis,  $\mathbb{Z}_2$  synchronization, spiked Wigner model, global convergence

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	This paper: randomly initialized AMP for $\mathbb{Z}_2$ synchronization	3
1.2	Main contributions and technical challenges	4
1.3	Notation	4
<b>2</b>	<b>Main results</b>	<b>5</b>
<b>3</b>	<b>Proof of Theorem 1</b>	<b>7</b>
3.1	Preliminaries	7
3.2	Non-asymptotic analysis for the AMP dynamics	9
<b>4</b>	<b>Discussions</b>	<b>19</b>
<b>A</b>	<b>Proof of auxiliary lemmas and claims</b>	<b>19</b>
A.1	Proof of Lemma 2	19
A.2	Proof of Lemma 3	22
A.3	Proof of Lemma 4	25
A.4	Proof of Claim (43)	32
A.5	Proof of Claim (53)	36
A.6	Proof of Claim (56)	39
A.7	Proof of Claim (59)	39
A.8	Proof of Lemma 5	40
A.9	Proof of inequality (70)	43

# 1 Introduction

The problem of estimating an unknown structured signal  $v^* \in \mathbb{R}^n$ , when given access to noisy observations

$$M = \lambda v^* v^{*\top} + W \in \mathbb{R}^{n \times n} \quad \text{with } \lambda > 0 \quad (1)$$

is of fundamental interest and has been investigated in a diverse array of contexts (Abbe et al., 2020; Candes and Plan, 2010; Chi et al., 2019; Johnstone, 2001; Keshavan et al., 2009; Singer, 2011; Zhong and Boumal, 2018). This model is commonly referred to as a deformed Wigner model or spiked Wigner model when the entries of the noise matrix  $W = [W_{ij}]_{1 \leq i,j \leq n}$  are independently drawn from Gaussian distributions — more precisely,  $W_{ii} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{2}{n})$  and  $W_{ji} = W_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n})$  for  $i \neq j$  — which serves as a prototypical model towards understanding the feasibility and fundamental limits of low-rank matrix estimation.

The spectral properties of the observed matrix  $M$  has been extensively studied (see, e.g. Baik et al. (2005); Capitaine et al. (2009); Cheng et al. (2021); Féral and Péché (2007); Péché (2006)), motivating the design of spectral methods when there is no structural information associated with  $v^*$  (Cai and Zhang, 2018; Chen et al., 2021; Keshavan et al., 2009; Singer, 2011; Yan et al., 2021). In practice, there is no shortage of applications where additional structural information about  $v^*$  is available *a priori*, examples including finite-group structure (Perry et al., 2018), cone constraints (Deshpande et al., 2014; Lesieur et al., 2017), sparsity (Berthet and Rigollet, 2013; Johnstone and Lu, 2009), among others. The presence of prior structure further exacerbates the nonconvexity issue when computing the maximum likelihood estimate or Bayes-optimal estimate, thereby presenting pressing needs for designing algorithms that can be executed efficiently. Remarkably, the approximate message passing (AMP) algorithm emerges as an efficient nonconvex paradigm that rises to the aforementioned challenge (Donoho et al., 2009; Feng et al., 2022). Originally proposed in the context of compressed sensing, AMP has served as not only a family of first-order iterative algorithms that enjoy rapid convergence (Bayati and Montanari, 2011a; Celentano and Montanari, 2022; Donoho et al., 2010; Fan, 2022; Rangan, 2011), but also a powerful machinery that assists in determining the performance of other statistical procedures in high-dimensional asymptotics (Bayati and Montanari, 2011b; Bu et al., 2020; Donoho and Montanari, 2016; Javanmard et al., 2016; Lelarge and Miolane, 2019; Li and Wei, 2021; Ma et al., 2018; Sur and Candès, 2019; Sur et al., 2019; Zhang et al., 2022).

Nevertheless, while the existing suite of AMP theory covers a wealth of applications, it remains inadequate in at least two aspects. To begin with, a dominant fraction of existing AMP theory is asymptotic in nature, in the sense that it predicts the AMP dynamics in the large- $n$  limit for any *fixed* iteration  $t$ . For this reason, prior AMP theory falls short of describing how AMP behaves after a growing number of iterations, which stands in contrast to other optimization-based procedures that often come with non-asymptotic analysis accommodating a large number of iterations (Chen and Wainwright, 2015; Chi et al., 2019; Keshavan et al., 2009; Ma et al., 2020). Another issue stems from the requirement of an informative initialization, that is, existing AMP theory for low-rank estimation often requires starting from a point that already enjoys non-vanishing correlation with the true signal (Celentano et al., 2021; Montanari and Venkataramanan, 2021; Zhong et al., 2021). While an informative initial estimate like spectral initialization is sometimes plausible and analyzable, this requirement presents a hurdle to understanding the effect of other widely adopted alternatives like random initialization. As shall be made clear shortly, tackling this issue might also necessitate a new non-asymptotic framework for AMP, due to the difficulty of tracking the AMP dynamics when the iterates exhibit only extremely weak correlation with the truth.

Inspired by the aforementioned issues, there has been growing interest in understanding the finite-sample performance of AMP. The first work of this kind was Rush and Venkataramanan (2018) (see also its follow-up work Cademartori and Rush (2023)), which studied AMP for sparse regression and permitted the total number of iterations to be as large as  $o(\frac{\log n}{\log \log n})$ . A recent work Li and Wei (2022) developed a non-asymptotic framework for the spiked Wigner models, which characterized the AMP behavior for up to  $O(\frac{n}{\text{poly}(\log n)})$  iterations. Although the theory therein is well-suited to spectrally initialized AMP, it remains in mystery whether randomly initialized AMP would be able to achieve the same performance as AMP with informative initialization.

## 1.1 This paper: randomly initialized AMP for $\mathbb{Z}_2$ synchronization

In this work, we attempt to address the above challenges by studying a concrete model called  $\mathbb{Z}_2$  synchronization. To be precise,  $\mathbb{Z}_2$  synchronization is a special case of the spiked Wigner model when the ground truth is known to have a discrete structure obeying  $v^* \in \{\pm \frac{1}{\sqrt{n}}\}^n$ . Here and throughout, we impose a prior distribution on  $v^* = [v_i^*]_{1 \leq i \leq n}$  such that

$$v_i^* \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\left(\pm \frac{1}{\sqrt{n}}\right), \quad 1 \leq i \leq n.$$

The goal is to reconstruct  $v^*$  on the basis of the measurements  $M$  (see (1)). This problem can be viewed as a basic example of a more general problem — synchronization over compact groups (Abbe et al., 2020; Chen and Candès, 2018; Gao and Zhang, 2022; Perry et al., 2018; Singer, 2011; Zhong and Boumal, 2018), and has an intimate connection to stochastic block models (Deshpande et al., 2017; Lelarge and Miolane, 2019).

**The AMP algorithm.** Due to the combinatorial nature of the underlying optimization problem, it is in general intractable to calculate the Bayes-optimal solution directly. This motivates the search for computationally feasible alternatives, for which AMP emerges as a natural and successful option (Celentano et al., 2021; Deshpande et al., 2017; Fan et al., 2021; Li and Wei, 2022). More concretely, given the initialization points  $x_0, x_1 \in \mathbb{R}^n$ , AMP tailored to  $\mathbb{Z}_2$  synchronization adopts the following update rule:

$$x_{t+1} = M\eta_t(x_t) - \langle \eta'_t(x_t) \rangle \eta_{t-1}(x_{t-1}), \quad t \geq 1, \tag{2}$$

where we denote  $\langle x \rangle := \frac{1}{n} \sum_{i=1}^n x_i$  for any vector  $x = [x_i]_{1 \leq i \leq n}$ , and the denoising function is given by<sup>1</sup>

$$\begin{aligned} \eta_t(x) &= \gamma_t \tanh(\pi_t x), \quad \text{for } t \geq 1 \\ \text{with } \pi_t &:= \sqrt{\max\{n(\|x_t\|_2^2 - 1), 1\}} \quad \text{and} \quad \gamma_t := \|\tanh(\pi_t x_t)\|_2^{-1}. \end{aligned} \tag{3}$$

Here, it is understood that  $\eta_t(\cdot)$ ,  $\eta'_t(\cdot)$  and  $\tanh(\cdot)$  are applied entrywise if the input argument is a vector.

Thus far, there are two analysis strategies that accommodate a growing number of iterations in the most challenging regime (i.e.  $\lambda > 1$ ). One attempt was made by Celentano et al. (2021), which proposed a three-stage hybrid algorithm that runs spectrally initialized AMP followed by natural gradient descent (NGD). It was conjectured therein that the third stage (i.e. NGD) is unnecessary. Recently, Li and Wei (2022) put forward another strategy to address this conjecture, showing that spectrally-initialized AMP alone is sufficient without the need of a third refinement stage. Despite the nonconvex nature of the underlying optimization problem, AMP with spectral initialization performs nearly the same as the Bayes-optimal estimate.

**Random initialization.** As alluded to previously, all existing AMP theories for this problem (Celentano et al., 2021; Li and Wei, 2022; Mondelli and Venkataraman, 2021; Montanari and Venkataraman, 2021) require informative initialization obtained by, for example, spectral methods. This leads to the following natural questions:

*Is spectral initialization necessary for the success of AMP?*

*Can we start with a simpler initialization scheme but still work equally well as spectral initialization?*

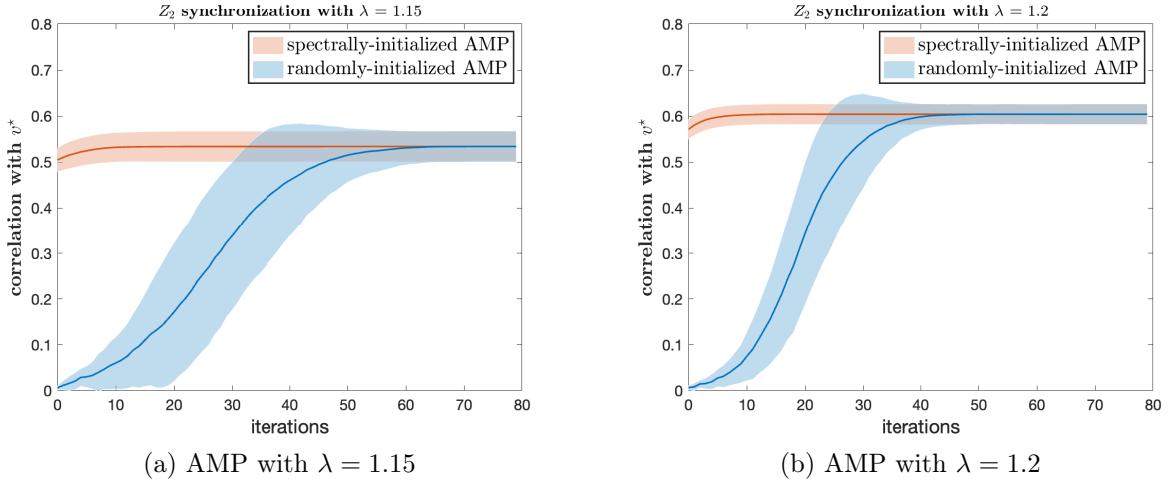
One initialization strategy that enjoys widespread adoption is to initialize AMP randomly; for instance,

$$x_1 \sim \mathcal{N}\left(0, \frac{1}{n}I_n\right) \quad (\text{independent of } M) \quad \text{and} \quad \eta_0(x_0) = 0. \tag{4}$$

In order to answer the above questions, let us first conduct a series of numerical experiments using (4), as reported in Figure 1. Encouragingly, AMP with random initialization seems to work surprisingly well: it only takes several tens of iterations to achieve nearly the same performance as spectrally initialized AMP (note that spectral initialization also consists of several tens of power iterations). Such encouraging numerical results motivate us to pursue in-depth theoretical understanding about the effect of random initialization upon AMP convergence, which was poorly understood in the literature.

---

<sup>1</sup>Note that for ease of analysis, we adopt a slightly different scaling from that of Deshpande et al. (2017), but they are equivalent up to global scaling.



**Figure 1.** The correlation of  $\eta_t(x_t)$  and  $v^*$  (i.e.  $\frac{|\langle \eta_t(x_t), v^* \rangle|}{\|\eta_t(x_t)\|_2}$ ) vs. iteration count  $t$  for AMP with both random and spectral initialization. Here,  $n = 10000$  and  $v_i^* \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\pm \frac{1}{\sqrt{n}})$  ( $1 \leq i \leq n$ ). We generate 20 independent copies of  $M$  according to (1) and report the averaged results, with the width of the shaded region reflecting (twice) the standard deviation. Figure (a) and (b) correspond to  $\lambda = 1.15$  and  $\lambda = 1.2$ , respectively.

## 1.2 Main contributions and technical challenges

In the present paper, we provide a non-asymptotic analysis that allows one to predict how AMP evolves over time from random initialization. Our theory is able to track the correlation of the AMP iterates and the truth  $v^*$ . In particular, we demonstrate in Theorem 1 that the signal component in the AMP iterates increases exponentially fast at the initial stage, taking no more than  $O(\frac{\log n}{\lambda-1})$  iterations to grow from  $\tilde{O}(\frac{1}{\sqrt{n}})$  to  $O(\sqrt{\lambda^2 - 1})$  (the latter of which coincides with the correlation of spectral initialization and the truth). Furthermore, once the signal component surpasses  $\frac{1}{2}\sqrt{\lambda^2 - 1}$  in magnitude, the finite-sample AMP dynamics are very well predicted by the state evolution recursion derived in the asymptotics. To the best of our knowledge, our paper delivers the first result to characterize the performance of AMP from random initialization, justifying and advocating the use of random initialization in practice.

Built upon the analysis recipe recently developed by Li and Wei (2022), the development of our theory requires novel ideas beyond this framework in order to track AMP from random initialization. Before continuing, we take a moment to single out the key challenges that need to be overcome.

- Prior theory based on state-evolution analysis falls short of providing “fine-grained” understanding about the iterates when they have vanishing correlation with the truth. More precisely, past theory fails to measure the progress of AMP during the initial stage when its signal component is  $o(1)$  (in fact, as small as  $\tilde{O}(\frac{1}{\sqrt{n}})$  when initialized), but instead treats the signal strength as 0 in the large- $n$  limit.
- Another technical challenge results from the complicated statistical dependency across iterations, which is particularly difficult to cope with when the algorithm starts with random initialization and when the number of iterations grows with the dimension  $n$ . While prior literature tackles this issue for other nonconvex optimization methods by resorting to either delicate leave-one-out decoupling arguments (see, e.g. Chen et al. (2019)) or global landscape analysis (see, e.g. Ge et al. (2017)), these approaches remain unavailable when analyzing AMP.

## 1.3 Notation

Throughout this paper, we use  $\varphi(\cdot)$  (resp.  $\varphi_n(\cdot)$ ) to denote the probability density function (p.d.f.) of a standard Gaussian random variable (resp. a Gaussian random vector  $\mathcal{N}(0, I_n)$ ). For any matrix  $M$ , we let  $\|M\|$  and  $\|M\|_{\text{F}}$  denote the spectral norm and the Frobenius norm of  $M$ , respectively. If not noted otherwise,

for any vector  $x \in [x_i]_{1 \leq i \leq n} = \mathbb{R}^n$ , we denote by  $|x|_{(i)}$  (resp.  $x_{(i)}$ ) the  $i$ -th largest absolute value (resp. value) of  $x$  in magnitude. We write  $\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$  as the unit sphere in  $\mathbb{R}^d$ . Moreover, for any two vectors  $x, y \in \mathbb{R}^n$ , we write  $x \circ y$  for their Kronecker product, namely,  $x \circ y = (x_1 y_1, \dots, x_n y_n)^\top \in \mathbb{R}^n$ . When a function is applied to a vector, it should be understood as being applied in a component-wise fashion; for instance, for any vector  $x = [x_i]_{1 \leq i \leq n}$ , we let  $x + 1 := [x_i + 1]_{1 \leq i \leq n}$ .

In addition, given two functions  $f(n)$  and  $g(n)$ , we write  $f(n) \lesssim g(n)$  or  $f(n) = O(g(n))$  to indicate that  $|f(n)| \leq c_1 g(n)$  for some universal constant  $c_1 > 0$  independent of  $n$ , and similarly,  $f(n) \gtrsim g(n)$  means that  $f(n) \geq c_2 |g(n)|$  for some universal constant  $c_2 > 0$ . We write  $f(n) = \tilde{O}(g(n))$  if  $f(n) = O(g(n))$  up to logarithm factors. We also adopt the notation  $f(n) \asymp g(n)$  to indicate that both  $f(n) \lesssim g(n)$  and  $f(n) \gtrsim g(n)$  hold simultaneously. Moreover, when we write  $f(n) \ll g(n)$  or  $f(n) = o(g(n))$ , it means  $f(n)/g(n) \rightarrow 0$  as  $n \rightarrow \infty$ ; we also write  $f(n) \gg g(n)$ , if  $g(n)/f(n) \rightarrow 0$  as  $n \rightarrow \infty$ . We use  $c, C$  to denote universal constants that do not depend on  $n$ ; note that constants may change from line to line.

## 2 Main results

In this section, we provide precise statements of our main theoretical guarantees for randomly initialized AMP. For notational convenience, let us introduce

$$\alpha_{t+1} := \lambda v^* \eta_t(x_t), \quad (5)$$

which captures the projection of the  $t$ -th iterate (after denoising) onto the direction of the truth  $v^*$ . In some sense, this quantity captures the size of the signal component carried by the  $t$ -th iterate. With this notation in place, we single out a key threshold as follows:

$$\varsigma := \min \left\{ t : \alpha_t \geq \frac{1}{2} \sqrt{\lambda^2 - 1} \right\}, \quad (6)$$

which reflects the time taken for the AMP iterate to carry a significant signal component (note that a random initial guess obeys  $|v^* \top x_1| \lesssim \tilde{O}\left(\frac{1}{\sqrt{n}}\right)$ , meaning that the initial signal component is exceedingly small). Additionally, we define the state-evolution recursion starting from the  $\varsigma$ -th iteration as follows:

$$\alpha_\varsigma^* = |\alpha_\varsigma| \quad \text{and} \quad \alpha_{t+1}^* = \lambda \left[ \int \tanh(\alpha_t^* (\alpha_t^* + x)) \varphi(dx) \right]^{1/2} \quad (\forall t \geq \varsigma). \quad (7)$$

Equipped with the above definitions, our main results are summarized in the following theorem.

**Theorem 1.** Consider the  $\mathbb{Z}_2$  synchronization problem with  $n^{-1/9} \log n \lesssim \lambda - 1 \leq 0.2$ . Suppose we run AMP (cf. (2) and (3)) with random initialization (4). Consider any  $t$  obeying  $1 \leq t \leq \frac{cn(\lambda-1)^5}{\log^2 n}$ , where  $c > 0$  is some universal constant. Then with probability at least  $1 - O(n^{-10})$ , the following results hold:

- (**Decomposition and error bound**) The AMP iterates admit the decomposition

$$x_t = \alpha_t v^* + \sum_{k=1}^{t-1} \beta_{t-k}^* \phi_k + \xi_{t-1}, \quad (8a)$$

where  $\alpha_t$  is defined in (5), the  $\phi_k$ 's are i.i.d. Gaussian vectors obeying  $\phi_k \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{n} I_n)$ , and

$$\|\eta_t(x_t)\|_2 = \|\beta_t\|_2 := \|(\beta_t^1, \beta_t^2, \dots, \beta_t^t)\|_2 = 1, \quad (8b)$$

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t \log n}{n(\lambda-1)^2}} + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}; \quad (8c)$$

- (**Crossing time**) The threshold  $\varsigma$  defined in (6) satisfies

$$\varsigma = O\left(\frac{\log n}{\lambda-1}\right); \quad (9)$$

- (*Non-asymptotic state evolution*) For any  $t$  obeying  $\varsigma \leq t \leq \frac{cn(\lambda-1)^5}{\log^2 n}$ , we have

$$\alpha_t^2 = \left( 1 + O\left( \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^5}} \right) \right) \alpha_t^{*2}, \quad (10)$$

where  $\{\alpha_t^*\}$  stand for the asymptotic state evolution parameters defined in (7).

**Remark 1** (Range of  $\lambda$ ). Theorem 1 only covers the regime where  $\lambda$  is larger than but close to 1. In fact,  $\lambda = 1$  represents the phase transition point for  $\mathbb{Z}_2$  synchronization (Deshpande et al., 2017), in the sense that (i) when  $\lambda < 1$ , no estimator performs better than the 0 estimator asymptotically, and (ii) when  $\lambda$  is strictly larger than 1, it is possible to achieve non-trivial correlation with  $v^*$ . We focus on the feasible regime by considering a more refined yet highly challenging case with  $\lambda - 1 \gtrsim n^{-1/9} \log n$  (so that  $\lambda$  can be very close to 1). While it is possible to improve the exponent 1/9, it is beyond the scope of this paper. The upper bound  $\lambda \leq 1.2$  is not crucial, and we make this assumption merely to simplify the presentation.

In the sequel, we provide some interpretations of Theorem 1 and discussions about its implications. It is assumed below that  $\lambda > 1$ .

**Gaussian approximation.** The first result (8a) in Theorem 1 asserts that each AMP iterate is composed of three components: (i) a signal component  $\alpha_t v^*$  that aligns with the true signal  $v^*$ , (ii) a noise component  $\sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k$  that is a linear combination of i.i.d. Gaussian vectors, and (iii) a residual component  $\xi_{t-1}$ . While this decomposition resembles that of Li and Wei (2022), we justify its validity even in the absence of carefully designed spectral initialization. A few remarks are in order.

- Regarding the noise component, the 1-Wasserstein distance between its distribution (denoted by  $\mu(\sum_{k=1}^t \beta_k^k \phi_k)$ ) and a Gaussian distribution  $\mathcal{N}(0, \frac{1}{n} I_n)$  is at most (Li and Wei, 2022, Lemma 9)

$$W_1\left(\mu\left(\sum_{k=1}^t \beta_k^k \phi_k\right), \mathcal{N}\left(0, \frac{1}{n} I_n\right)\right) \lesssim \sqrt{\frac{t \log n}{n}}. \quad (11)$$

For  $t$  not too large, the noise component well approximates a Gaussian vector  $\mathcal{N}(0, \frac{1}{n} I_n)$ .

- Regarding the signal component  $\alpha_t v^*$ , it is self-evident that  $\alpha_t$  governs how effective AMP is in recovering the true signal. Importantly, once  $\alpha_t$  exceeds the threshold  $\frac{1}{2} \sqrt{\lambda^2 - 1}$ , it follows a non-asymptotic state evolution that closely resembles the asymptotic counterpart  $\alpha_t^*$  (see (10)), a result that is made possible thanks to the non-asymptotic nature of our analysis.

To summarize, up to a small error term at most  $\tilde{O}\left(\sqrt{\frac{t}{n(\lambda-1)^2}} + \sqrt{\frac{1}{n(\lambda-1)^3}}\right)$ , the AMP iterate is approximately

$$x_t \approx \alpha_t v^* + \mathcal{N}\left(0, \frac{1}{n} I_n\right), \quad t < O\left(\frac{n(\lambda-1)^5}{\log^2 n}\right)$$

even when initialized randomly. An asymptotic version of this observation has been made in Montanari and Venkataramanan (2021), although the result therein required both informative initialization and a fixed  $t$  that does not grow with  $n$ .

**Dynamics after random initialization.** Compared to Li and Wei (2022, Theorem 4), the most challenging element of Theorem 1 lies in analyzing the initial stage after random initialization. As shall be made clear from our analysis, we can understand the AMP trajectory by dividing it into three phases.

- **Phase #1: escape from random initialization.** When initialized randomly with  $x_1 \sim \mathcal{N}(0, \frac{1}{n} I_n)$ , AMP starts with an extremely small signal component about the order of  $\tilde{O}(\frac{1}{\sqrt{n}})$ , for which the canonical state evolution becomes vacuous. To overcome this technical hurdle, we develop fine-grained characterizations regarding how  $\alpha_t$  evolves in this phase (before  $|\alpha_t|$  surpasses  $\sqrt{\lambda-1} n^{-1/4}$ ), that is,

$$\alpha_{t+1} \approx \lambda \alpha_t + \lambda g_{t-1}, \quad \text{with } g_{t-1} \sim \mathcal{N}\left(0, \frac{1}{n}\right); \quad (12)$$

see Section A.4 for details. From this approximate noisy recursion, while the signal component might be initially buried under the noise term, it only takes  $O(\frac{\log n}{\lambda-1})$  iterations for the signal component to rise above the noise size and reach the order of  $\sqrt{\lambda-1}n^{-1/4}$  (see Section 3.2.2).

- **Phase #2: exponential growth.** Once the signal component exceeds  $\sqrt{\lambda-1}n^{-1/4}$  in size, the AMP iterate correlates non-trivially with the true signal. Interestingly, the signal strength  $\alpha_t$  starts to grow exponentially until reaching the order of  $\sqrt{\lambda^2-1}$ . As we shall justify in Section 3.2.2,  $\alpha_{t+1}$  obeys

$$\alpha_{t+1} \geq \sqrt{1 + \frac{1-o(1)}{3}(\lambda-1)} \alpha_t \quad (13)$$

in this phase, which accounts for at most  $O(\frac{\log n}{\lambda-1})$  iterations.

- **Phase #3: local refinement.** Upon reaching the order of  $\sqrt{\lambda^2-1}$ ,  $\alpha_t$  enters a local refinement phase, during which randomly initialized AMP behaves similarly as AMP with spectral or other informative initialization. In this phase, the asymptotic state evolution (7) also starts to be effective when predicting the evolution of  $\alpha_t$  (see (10)). As we shall show in Section 3.2.4, the signal strength  $\alpha_t$  satisfies

$$|\alpha_{t+1}^2 - \alpha^{*2}| \lesssim (1 - (\lambda - 1))^{t-\varsigma} + \tilde{O}\left(\sqrt{\frac{t + \frac{1}{\lambda-1}}{n(\lambda-1)^5}}\right), \quad (14)$$

where  $\alpha^*$  (determined by  $\lambda$ ) denotes the limit of  $\alpha_t^*$  (cf. (7)) and is unique solution of

$$\alpha^{*2} = \lambda^2 \mathbb{E}[\tanh(\alpha^*(\alpha^* + G))], \quad \text{for } G \sim \mathcal{N}(0, 1). \quad (15)$$

**Asymptotic optimality.** As was shown previously (see e.g. Celentano et al. (2021, Lemma A.7) and Li and Wei (2022, Remark 4)), we can construct an AMP-based estimator whose risk coincides with that of the Bayes-optimal estimator  $\hat{X}^{\text{bayes}} := \mathbb{E}[v^* v^{*\top} | M]$ . More precisely, taking the AMP-based estimator as

$$u_t := \frac{1}{\lambda \sqrt{n(\alpha_t^2 + 1)}} \tanh(\pi_t x_t), \quad (16)$$

its asymptotic risk satisfies (see Section B and Deshpande et al. (2017)):

$$\lim_{t \rightarrow \infty} \lim_{n \rightarrow \infty} \mathbb{E}[\|v^* v^{*\top} - u_t u_t^\top\|_F^2] = \lim_{n \rightarrow \infty} \mathbb{E}[\|v^* v^{*\top} - \hat{X}^{\text{bayes}}\|_F^2] = 1 - \frac{\alpha^{*4}}{\lambda^4}, \quad (17)$$

where  $\alpha^*$  is the fixed point of the limiting state evolution (cf. (15)). This together with the non-asymptotic results in Theorem 1 leads to a more refined risk characterization, as we shall prove in Section B.

**Corollary 1.** *With probability at least  $1 - O(n^{-10})$ , there exists some  $t = O(\frac{\log n}{\lambda-1})$  such that*

$$\|v^* v^{*\top} - u_t u_t^\top\|_F^2 = 1 - \frac{\alpha^{*2}}{\lambda^4} + O\left(\sqrt{\frac{\log^4 n}{n(\lambda-1)^6}}\right). \quad (18)$$

### 3 Proof of Theorem 1

In this section, we present the proof of our main result: Theorem 1. We find it helpful to introduce the following notation that helps streamline the presentation:

$$v_t := \alpha_t v^* + \sum_{k=1}^{t-1} \beta_{t-k}^k \phi_k. \quad (19)$$

#### 3.1 Preliminaries

Before we embark on our proof of the main theorem, we collect a couple of useful results that shall be used frequently throughout this proof.

**Concentration results.** We first record several useful concentration results from [Li and Wei \(2022\)](#). Here and throughout, we let  $|x|_{(i)}$  denote the magnitude of the  $i$ -th largest entry (in magnitude) of  $x \in \mathbb{R}^n$ .

**Lemma 1.** Consider a collection of random vectors  $\{\phi_k\}_{1 \leq k < t}$  in  $\mathbb{R}^n$ . Suppose that for each  $1 \leq k \leq t-1 < n$ ,  $\phi_k = (\phi_{k,1}, \phi_{k,2}, \dots, \phi_{k,n})$  is i.i.d. drawn from  $\mathcal{N}(0, \frac{1}{n}I_n)$ . Consider the following set

$$\begin{aligned} \mathcal{E}_s := & \left\{ \{\psi_k\}_{k=1}^{t-1} : \max_{1 \leq k \leq t-1} \|\psi_k\|_2 < 1 + C\sqrt{\frac{\log \frac{n}{\delta}}{n}} \right\} \cap \left\{ \{\psi_k\}_{k=1}^{t-1} : \sup_{a \in \mathcal{S}^{t-2}} \left\| \sum_{k=1}^{t-1} a_k \psi_k \right\|_2 < 1 + C\sqrt{\frac{t \log \frac{n}{\delta}}{n}} \right\} \\ & \cap \left\{ \{\psi_k\}_{k=1}^{t-1} : \sup_{a \in \mathcal{S}^{t-2}} \sum_{i=1}^s \left| \sum_{k=1}^{t-1} a_k \psi_k \right|_{(i)}^2 < \frac{C(t+s) \log \frac{n}{\delta}}{n} \right\} \cap \left\{ \{\psi_k\}_{k=1}^{t-1} : \max_{1 \leq k < t, 1 \leq i \leq n} |\psi_{k,i}| < C\sqrt{\frac{\log \frac{n}{\delta}}{n}} \right\}, \end{aligned}$$

and denote  $\mathcal{E} := \bigcap_{s=1}^n \mathcal{E}_s$ . Then there exists some large enough constant  $C > 0$  such that, for every  $\delta > 0$ ,

$$\mathbb{P}(\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}) \geq 1 - \delta.$$

In particular, by setting  $\delta = n^{-11}$ , we see that the following event happens with probability at least  $1 - O(n^{-11})$ :

$$\left| \max_{1 \leq k \leq t-1} \|\phi_k\|_2 - 1 \right| \lesssim \sqrt{\frac{\log n}{n}}, \quad \text{and} \quad \max_{1 \leq k \leq t, 1 \leq i \leq n} |\phi_{k,i}| \lesssim \sqrt{\frac{\log n}{n}}.$$

This lemma is a consequence of standard concentration of measure for Gaussian random vectors ([Massart, 2007](#)); its proof can be found in [Li and Wei \(2022, Section D.1.1\)](#) and is hence omitted for brevity.

**Properties of  $\eta_t$ ,  $\pi_t$ , and  $\gamma_t$  (cf. (3)).** Next, we summarize several basic properties about the three sets of key quantities defined in (3). We begin by gathering several basic properties for our choices of  $\pi_t$  and  $\gamma_t$  defined in (3); the proof is deferred to Section A.1.

**Lemma 2.** Suppose the decomposition (8a) is valid with  $\|\xi_{t-1}\|_2 \lesssim 1$ . With probability at least  $1 - O(n^{-10})$ , the following properties hold true:

$$\frac{1}{\sqrt{n}} \pi_t = |\alpha_t| + O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2} \wedge \frac{1}{|\alpha_t|} \left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)\right); \quad (20a)$$

$$\gamma_t^{-2} = n \int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(dx) + \pi_t^2 O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right). \quad (20b)$$

Additionally, one has

$$\int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(dx) = \frac{\pi_t^2}{n}(\alpha_t^2 + 1) + O\left(\frac{\pi_t^4}{n^2}\right), \quad (20c)$$

which in turn implies that

$$\gamma_t^{-2} = \pi_t^2 \left( \alpha_t^2 + 1 + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right) \right). \quad (20d)$$

Next, let us single out several properties about the quantity  $\eta_t$  in the lemma below, whose proof is postponed to Section A.2.

**Lemma 3.** Consider any  $1 \leq t \leq n$  and suppose that  $\|\xi_{t-1}\|_2$  satisfies

$$\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}} \quad \text{for all } t \lesssim \frac{\log n}{\lambda - 1}. \quad (21)$$

Then the following properties hold true with probability at least  $1 - O(n^{-11})$ :

- If  $t \lesssim \frac{n(\lambda-1)^4}{\log^2 n}$  and  $\|\xi_{t-1}\|_2 \lesssim \frac{1}{\sqrt{\log n}}$ , then any  $x \in \mathbb{R}$  obeys

$$|\eta_t(x)| \lesssim |x|, \quad |\eta'_t(x)| \lesssim 1 =: \rho, \quad |\eta''_t(x)| \lesssim \sqrt{n} =: \rho_1, \quad |\eta_t^{(m)}(x)| \lesssim n =: \rho_2; \quad (22a)$$

- If  $t \lesssim \frac{\log n}{\lambda-1}$  and  $\alpha_t \lesssim \sqrt{\lambda-1}n^{-0.1}$ , then one has  $\pi_t \lesssim \sqrt{\lambda-1}n^{0.4}$ ; and for any  $x$  obeying  $|x| \lesssim \sqrt{\log n/n}$ , we have

$$\eta'_t(x) = 1 + O((\lambda-1)n^{-0.2} \log n) \quad \text{and} \quad |\eta''_t(x)| \lesssim (\lambda-1)n^{0.8}|x|; \quad (22b)$$

- If  $t \lesssim \frac{\log n}{\lambda-1}$  and  $\alpha_t \lesssim \sqrt{\lambda-1}n^{-1/4}$ , then one has  $\pi_t \lesssim (\lambda-1)^{-3/4}n^{1/4} \log n$ ; for any  $x$ , one can find a quantity  $c_0 \lesssim (\log^2 n)/\sqrt{n(\lambda-1)^3}$  independent from  $x$ , and another quantity  $|c_x| \lesssim n|x|^5(\log^4 n)/(\lambda-1)^3$  depending on  $x$ , such that

$$\eta_t(x) = (1 - c_0) \left( x - \frac{1}{3} \pi_t^2 x^3 + c_x \right). \quad (22c)$$

### 3.2 Non-asymptotic analysis for the AMP dynamics

We are now in a position to present the proof of our main theorem. The structure of our proof is outlined in what follows.

- Firstly, focusing on the initial stage obeying  $t \leq \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ , we develop an upper bound on  $\|\xi_t\|_2$  in Section 3.2.1 as follows:

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}; \quad (23)$$

here,  $\varsigma$  is a threshold defined in (6) and  $c > 0$  is some constant small enough. This is accomplished by means of an inductive argument.

- Secondly, we investigate in Section 3.2.2 how the signal strength  $\alpha_t$  evolves during the execution of AMP. Crucially, recalling that  $\varsigma$  reflects the first time  $t$  that satisfies  $|\alpha_t| \gtrsim \sqrt{\lambda^2 - 1}$  (cf. (6)), we demonstrate that

$$\varsigma \lesssim \frac{\log n}{\lambda-1}; \quad (24)$$

in words, in spite of random (and hence uninformative) initialization, it takes AMP at most  $O(\frac{\log n}{\lambda-1})$  iterations to find an informative estimate.

- Thirdly, with the above control of  $\varsigma$  in place, we go on to develop a more complete upper bound on  $\|\xi_t\|_2$  that covers the iterations after  $\varsigma$ , that is,

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{t \mathbf{1}(t > \varsigma) \log n}{n(\lambda-1)^2}} + \sqrt{\frac{\min\{t, \varsigma\}^3 \log n}{n}} \quad (25)$$

for any  $t < \frac{cn(\lambda-1)^5}{\log^2 n}$ . This is the main content of Section 3.2.3, accomplished again via an inductive argument.

- Finally, after the iteration number exceeds the threshold  $\varsigma$ , we demonstrate in Section 3.2.4 that the asymptotic state evolution (the one characterizing large-system limits) becomes fairly accurate in the finite-sample/finite-time regime. In particular, an intimate connection is established between the non-asymptotic state evolution and its asymptotic analog, which plays a critical role in characterizing the finite-sample convergence behavior of AMP.

These four steps will be explained in detail in the sequel.

### 3.2.1 Controlling $\xi_t$ when $t \leq \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ (Proof of Claim (23))

In this subsection, we establish the claimed bound (23) for  $\|\xi_t\|_2$ , which leverages on ideas from Li and Wei (2022). To begin with, let us restate Li and Wei (2022, Theorem 2) below, with slight simplification tailored to  $\mathbb{Z}_2$  synchronization (i.e., through the use of the properties  $\|\beta_t\|_2^2 = 1$ ,  $E_t = 0$ , and (22a)). For notational convenience, define  $\kappa_t > 0$  such that

$$\begin{aligned} \kappa_t^2 := \max \left\{ \left\langle \int \left[ x \eta'_t \left( \alpha_t v^* + \frac{1}{\sqrt{n}} x \right) - \frac{1}{\sqrt{n}} \eta''_t \left( \alpha_t v^* + \frac{1}{\sqrt{n}} x \right) \right]^2 \varphi_n(dx) \right\rangle, \right. \\ \left. \left\langle \int \left[ \eta'_t \left( \alpha_t v^* + \frac{1}{\sqrt{n}} x \right) \right]^2 \varphi_n(dx) \right\rangle \right\}, \end{aligned} \quad (26)$$

where we recall that  $\varphi_n(\cdot)$  is the p.d.f. of  $\mathcal{N}(0, I_n)$  and for any vector  $x = [x_i]_{1 \leq i \leq n}$ , we denote  $\langle x \rangle := \frac{1}{n} \sum_{i=1}^n x_i$  and  $x^2 = [x_i^2]_{1 \leq i \leq n}$ . We shall work with the following assumptions.

**Assumption 1.** For any  $1 \leq t \leq n$ , consider arbitrary vectors  $\mu_t \in \mathcal{S}^{t-1}$ ,  $\xi_{t-1} \in \mathbb{R}^n$ , and coefficients  $(\alpha_t, \beta_{t-1}) \in \mathbb{R} \times \mathbb{R}^{t-1}$  that might all be statistically dependent on  $\phi_k$ . Let  $v_t$  be defined as in (19). We assume the existence of (possibly random) quantities  $A_t, B_t, D_t$  such that with probability at least  $1 - O(n^{-11})$ , the following inequalities hold:

$$\left| \sum_{k=1}^{t-1} \mu_t^k \left[ \langle \phi_k, \eta_t(v_t) \rangle - \langle \eta'_t(v_t) \rangle \beta_{t-1}^k \right] \right| \leq A_t, \quad (27a)$$

$$\left| v^{*\top} \eta_t(v_t) - v^{*\top} \int \eta_t \left( \alpha_t v^* + \frac{\|\beta_{t-1}\|_2}{\sqrt{n}} x \right) \varphi_n(dx) \right| \leq B_t, \quad (27b)$$

$$\left\| \sum_{k=1}^{t-1} \mu_t^k \phi_k \circ \eta'_t(v_t) - \frac{1}{n} \sum_{k=1}^{t-1} \mu_t^k \beta_{t-1}^k \eta''_t(v_t) \right\|_2^2 - \kappa_t^2 \leq D_t. \quad (27c)$$

Under these assumptions, Li and Wei (2022, Theorem 2) developed a general non-asymptotic characterization for AMP iterates as follows.

**Theorem 2.** [Adapted from Li and Wei (2022, Theorem 2)] Suppose that Assumption 1 holds, and consider any  $t \leq n$ . With probability at least  $1 - O(n^{-11})$ , the AMP iterates (2) for  $\mathbb{Z}_2$ -synchronization satisfy the decomposition (8a) with  $\|\beta_t\|_2^2 = 1$  and

$$\alpha_{t+1} = \lambda v^{*\top} \int \eta_t \left( \alpha_t v^* + \frac{1}{\sqrt{n}} x \right) \varphi_n(dx) + \Delta_{\alpha,t} \quad (28)$$

where the residual terms obey

$$|\Delta_{\alpha,t}| \lesssim B_t + |v^{*\top} \eta_t(x_t) - v^{*\top} \eta_t(v_t)| \lesssim B_t + \|\xi_{t-1}\|_2, \quad (29a)$$

$$\|\xi_t\|_2 \leq \sqrt{\kappa_t^2 + D_t} \|\xi_{t-1}\|_2 + O \left( \sqrt{\frac{t \log n}{n}} + A_t + \sqrt{(1 + t \mathbf{1}_{t \leq \varsigma}) \log n} \|\xi_{t-1}\|_2^2 + \sqrt{\frac{t \log n}{n}} \|\xi_{t-1}\|_2 \right). \quad (29b)$$

**Remark 2.** With regards to the above bound (29b) for  $\|\xi_t\|_2$ , a direct application of Li and Wei (2022, Theorem 2) results in a term  $\sqrt{t \log n} \|\xi_{t-1}\|_2^2$  (as opposed to  $\sqrt{(1 + t \mathbf{1}_{t \leq \varsigma}) \log n} \|\xi_{t-1}\|_2^2$  in (29b)). We make slight modifications here to make it better-suited for the current setting.

- (i) When  $t \leq \varsigma$ , such a term  $\sqrt{t \log n} \|\xi_{t-1}\|_2^2$  works fine for our purpose;
- (ii) When  $t > \varsigma$  (so that  $\alpha_t$  exceeds the order of  $\sqrt{\lambda^2 - 1}$ ), one can simply invoke Li and Wei (2022, display (249)) to improve the factor in front of  $\|\xi_{t-1}\|_2^2$  from  $\sqrt{t \log n}$  to  $\sqrt{\log n}$ .

Putting these together leads to the claimed bound (29b). Notably, this seemingly minor change turns out to be essential in order to push the number of iterations to  $O(n/\text{poly}(\log n))$  instead of  $O(\sqrt{n}/\text{poly}(\log n))$ .

With Theorem 2 in mind, in order to control  $|\Delta_{\alpha,t}|$  and  $\|\xi_t\|_2$ , it boils down to determining  $A_t, B_t, D_t$ , and  $\kappa_t$ , respectively.

- **Bounding  $A_t, B_t, D_t$ .** Repeating the same analysis as in [Li and Wei \(2022, Section D.2\)](#), we obtain

$$A_t \lesssim \sqrt{\frac{t \log n}{n}}, \quad B_t \lesssim \sqrt{\frac{t \log n}{n}}, \quad D_t \lesssim \sqrt{\frac{t \log^2 n}{n}}. \quad (30)$$

The only term that needs more discussion is  $A_t$ , as [Li and Wei \(2022\)](#) only proved that  $A_t \lesssim \frac{1}{\alpha_t} \sqrt{\frac{t \log n}{n}}$  (taking  $s = 1$  therein) for AMP with independent initialization. To get rid of the prefactor  $1/\alpha_t$ , we rely on an improved control of  $\eta_t(x)$  (cf. (22a)). In particular, property (22a) tells us that

$$\|\eta_t(v_t)\|_2 \lesssim \|v_t\|_2 = \left\| \alpha_t v^\star + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_2 \lesssim 1,$$

where the last inequality can be found in display (87). In turn, this leads to

$$\|\nabla_\Phi f_\theta(\Phi)\|_2 \lesssim \frac{1}{\sqrt{n}} \quad (31)$$

through the same analyses as detailed around [Li and Wei \(2022, Section D.2.1, inequality \(229\)\)](#). Here,  $\nabla_\Phi f_\theta(\Phi)$  is the key quantity to control  $A_t$  in [Li and Wei \(2022, Section D.2.1\)](#), and our desired bound for  $A_t$  follows immediately. Given that this only consists of very minor and straightforward changes to [Li and Wei \(2022, Section D.2.1\)](#), we omit the details for brevity and refer the readers to [Li and Wei \(2022, Section D.2.1\)](#) for more details.

- **Bounding  $\kappa_t$ .** The main step then comes down to bounding  $\kappa_t$ . Towards this end, we claim that the following relation holds for  $\kappa_t$ , whose proof is postponed to Section A.3.

**Lemma 4.** *With probability at least  $1 - O(n^{-10})$ , the following results hold true:*

- Under the inductive assumption (23) for  $\xi_{t-1}$ , one has

$$\kappa_t \leq 1 + o\left(\frac{\lambda - 1}{\log n}\right) \quad (32a)$$

provided that  $t \leq \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ ;

- Under the inductive assumption (25) for  $\xi_{t-1}$ , one has

$$\kappa_t \leq 1 - \frac{1}{15}(\lambda - 1), \quad (32b)$$

provided that  $t \leq \frac{cn(\lambda-1)^5}{\log^2 n}$  and  $|\alpha_t| \gtrsim \sqrt{\lambda^2 - 1}$ .

With the above estimates of  $A_t, B_t, D_t, \kappa_t$  in place, we are ready to apply Theorem 2. Under the inductive assumption (23), the recursive formula (29) in Theorem 2 taken together with (30) yields

$$\|\xi_t\|_2 \leq \sqrt{\kappa_t + \sqrt{\frac{t \log^2 n}{n}}} \|\xi_{t-1}\|_2 + O\left(\sqrt{\frac{t \log n}{n}} + \sqrt{(1 + t \mathbf{1}_{t \leq \varsigma}) \log n} \|\xi_{t-1}\|_2^2 + \sqrt{\frac{t \log n}{n}} \|\xi_{t-1}\|_2\right), \quad (33)$$

which combined with (32) further implies that

$$\|\xi_t\|_2 \leq \left(1 + o\left(\frac{\lambda - 1}{\log n}\right) + O\left(\sqrt{\frac{t^4 \log^2 n}{n}}\right)\right) \|\xi_{t-1}\|_2 + O\left(\sqrt{\frac{t \log n}{n}}\right), \quad (34)$$

with the proviso that  $t \leq \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ .

We are now ready to prove relation (23) via induction. To verify its validity for the base case (i.e.  $t = 1$ ), we note that by construction (see, e.g. [\(Li and Wei, 2022, Step 3, Proof of Theorem 1\)](#)),  $\xi_1$  takes the form

$$\xi_1 = \left( \frac{\sqrt{2}}{2} - 1 \right) z_1 z_1^\top W z_1, \quad \text{where } z_1 = \eta_1(x_1) \text{ is independent of } W.$$

Elementary calculations reveal that, with probability at least  $1 - O(n^{-11})$ ,

$$\|\xi_1\|_2 = \left| \frac{\sqrt{2}}{2} - 1 \right| \cdot \|z_1\|_2 \cdot |z_1^\top W z_1| \lesssim \sqrt{\frac{\log n}{n}}, \quad (35)$$

given that  $\|z_1\|_2 = 1$  and  $z_1^\top W z_1 \sim \mathcal{N}(0, \frac{2}{n} I_n)$ . This already establishes (23) for the base case with  $t = 1$ . Next, consider the case where  $t \leq \varsigma \wedge \frac{\log n}{c(\lambda-1)}$  for some small enough constant  $c > 0$ . Given that  $\sqrt{\frac{t^4 \log^2 n}{n}} = o(\frac{\lambda-1}{\log n})$  under our assumption on  $\lambda - 1$ , the recursive relation (34) immediately leads to

$$\begin{aligned} \|\xi_t\|_2 &\leq \left( 1 + o\left(\frac{\lambda-1}{\log n}\right) \right) \|\xi_{t-1}\|_2 + O\left(\sqrt{\frac{t \log n}{n}}\right) \\ &\leq \left( 1 + o\left(\frac{\lambda-1}{\log n}\right) \right)^{t-1} \|\xi_1\|_2 + \sum_{j=0}^{t-2} \left( 1 + o\left(\frac{\lambda-1}{\log n}\right) \right)^j O\left(\sqrt{\frac{(t-j) \log n}{n}}\right) \\ &\lesssim \sqrt{\frac{t^3 \log n}{n}} \end{aligned} \quad (36)$$

for all  $t \leq \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ , as claimed.

**Remark 3.** Careful readers might note that the recursive formula established in (34) for  $t \leq \varsigma \wedge \frac{\log n}{c(\lambda-1)}$  does not rely on the relation (24) (a relation that shall be established in the next subsection).

### 3.2.2 Evolution of $\alpha_t$ and a bound on $\varsigma$ (Proof of Claim (24))

We now move on to establish the claim (24) concerning an upper bound on the threshold  $\varsigma$ , which requires careful analysis about how the signal strength  $\alpha_t$  evolves at the initial stage. Towards this end, we divide into two cases based on the magnitude of  $\alpha_t$ , which we shall detail after presenting several preliminary facts.

**Preliminary facts.** Before proceeding, we first recall some additional preliminary facts already established in [Li and Wei \(2022\)](#). From the analysis of [Li and Wei \(2022, Theorem 1\)](#), we know that: by construction,

$$\xi_t \in \text{span}(U_{t-1}) = \text{span}\left\{\eta_1(x_1), \dots, \eta_{t-1}(x_{t-1})\right\}, \quad (37)$$

where  $U_{t-1} \in \mathbb{R}^{n \times (t-1)}$  is a matrix whose columns are formed by a set of orthonormal basis  $\{z_1, \dots, z_{t-1}\}$ . In fact, we can specify  $U_t$  in a more explicit manner. Following [Li and Wei \(2022, Section 4.1\)](#), let us define

$$z_1 := \frac{\eta_1(x_1)}{\|\eta_1(x_1)\|_2} \in \mathbb{R}^n \quad \text{and} \quad W_1 := W \in \mathbb{R}^{n \times n}, \quad (38a)$$

which are statistically independent from each other; and then any  $2 \leq t \leq n$ , we can define the following objects recursively:

$$U_{t-1} := [z_k]_{1 \leq k \leq t-1} \in \mathbb{R}^{n \times (t-1)}, \quad (38b)$$

and also

$$z_t := \frac{(I_n - U_{t-1} U_{t-1}^\top) \eta_t(x_t)}{\|(I_n - U_{t-1} U_{t-1}^\top) \eta_t(x_t)\|_2}, \quad (38c)$$

$$W_t := (I_n - z_{t-1} z_{t-1}^\top) W_{t-1} (I_n - z_{t-1} z_{t-1}^\top), \quad (38d)$$

where  $\{x_t\}$  is the sequence generated by the AMP updates (2). This process thus leads to more explicit forms for  $\{U_t\}$  and the orthonormal basis  $\{z_t\}$  (see Li and Wei (2022, Section 4.1) for the orthonormality of  $\{z_t\}$ ). What is more, the orthonormality of  $\{z_t\}$  reveals the decomposition

$$\eta_t(x_t) = \sum_{k=1}^t \beta_t^k z_k, \quad \text{with } \beta_t^k := \langle \eta_t(x_t), z_k \rangle, \quad (39)$$

which satisfies  $\|\eta_t(x_t)\|_2 = \|\beta_t\|_2$  with  $\beta_t = [\beta_t^1, \dots, \beta_t^t]$ . Additionally, we find it convenient to generate

$$\phi_k := W_k z_k + \zeta_k, \quad \text{where } \zeta_k := \left( \frac{\sqrt{2}}{2} - 1 \right) z_k^\top W_k z_k \cdot z_k + \sum_{i=1}^{k-1} g_i^k z_i, \quad 1 \leq k \leq n, \quad (40)$$

where the  $g_i^k$ 's are independently drawn from  $\mathcal{N}(0, \frac{1}{n} I_n)$ . The following properties have been shown in Li and Wei (2022, Lemma 2), which play a crucial role in our subsequent analysis:

- $\phi_k \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{n} I_n)$ , for  $1 \leq k \leq n$ ;
- The randomness of  $\phi_k$  only comes from  $W_k$ , and  $\phi_k$  is independent of  $x_1$  and  $\{z_i\}_{i < k}$ .
- $x_k$  and  $z_k$  are conditionally independent from  $W_k$  given  $\{z_i\}_{i < k}$  and  $x_1$ .
- $\phi_k$  is independent from  $\{x_j\}_{j \leq k}$  and  $\{z_j\}_{j \leq k}$ .

**Stage I: small correlation** ( $|\alpha_t| \lesssim \sqrt{\lambda-1} n^{-1/4}$ ). Let us start from the very beginning when the correlation coefficient  $\alpha_t$  is reasonably small. Towards this, we define a threshold  $\tau_0$  such that

$$\tau_0 := \max \{ \tau : |\alpha_t| \lesssim \sqrt{\lambda-1} n^{-1/4} \text{ for all } t \leq \tau \}; \quad (41)$$

in words,  $\tau_0 + 1$  represents the first term that exceeds the level of  $\sqrt{\lambda-1} n^{-1/4}$ . In the following, we would like to prove that, with probability at least  $1 - O(n^{-10})$ , this threshold is not too large in the sense that

$$\tau_0 \lesssim \frac{\log n}{\lambda-1}. \quad (42)$$

*Proof of Claim (42).* In order to establish this result (42), we first state an important claim: the AMP iterates — when initialized at a random point — satisfy the following recursive relation with high probability:

$$\alpha_{t+1} = \lambda^{t-k+1} \alpha_k + \sum_{i=1}^{t-k+1} \lambda^i g_{t-i} + O\left(\lambda^{t-k} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}}\right) \quad (43)$$

for any  $1 \leq k \leq t$ , where we denote

$$g_k := v^* \top \phi_k \quad (1 \leq k \leq t) \quad \text{and} \quad g_0 = 0. \quad (44)$$

This claimed relation lies at the heart of the analysis for Stage I, in which the correlation between the AMP iterate and  $v^*$  keeps growing to a non-trivial value. To streamline the presentation, we defer the proof of this claim to Section A.4.

Equipped with the above recursive formula (43), we now turn to proving the relation (42). Define  $t_i := C' i \log n$  for some quantity  $C' = \frac{C''}{\lambda-1}$ , where  $C''$  is some large enough constant. Observe that

$$\begin{aligned} \mathbb{P}\left(|\alpha_k| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, \text{ for all } k \leq 201C' \log n\right) &\leq \mathbb{P}\left(|\alpha_{t_i+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, \text{ for all } 1 \leq i \leq 200\right) \\ &= \prod_{i=1}^{200} \mathbb{P}\left(|\alpha_{t_i+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} \mid |\alpha_{t_j+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, \forall 1 \leq j < i\right). \end{aligned}$$

To control the right-hand side of the above relation, consider the following random variable

$$X_i := \sum_{j=1}^{C' \log n} \lambda^j g_{t_i-j} \sim \mathcal{N} \left( 0, \frac{\lambda^{2C' \log n + 2} - \lambda^2}{n(\lambda^2 - 1)} \right) \quad (45)$$

for each  $1 \leq i \leq 200$ . Armed with this piece of notation, invoking relation (43) gives

$$\alpha_{t_i+1} = \lambda^{t_i-t_{i-1}} \alpha_{t_{i-1}+1} + X_i + O \left( \lambda^{t_i-t_{i-1}-1} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}} \right). \quad (46)$$

As mentioned in the above preliminary facts, each  $\phi_j$  is independent with the AMP iterate  $x_i$  for  $i \leq j$  and therefore  $\alpha_{i+1}$ , given that  $\alpha_{i+1} = v^* \eta_i(x_i)$ . As a result, the random variable  $X_i$  defined above is independent from  $\alpha_{t_j+1}$  for all  $j \leq i-1$ . Taking this together with the relation (43) then leads to

$$\begin{aligned} & \mathbb{P} \left( |\alpha_{t_i+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} \mid |\alpha_{t_j+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, 1 \leq j < i \right) \\ & \leq \mathbb{P} \left( |\lambda^{t_i-t_{i-1}} \alpha_{t_{i-1}+1} + X_i| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \lambda^{t_i-t_{i-1}-1} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}} \mid |\alpha_{t_j+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, 1 \leq j < i \right) \\ & = \mathbb{P} \left( |\lambda^{t_i-t_{i-1}} \alpha_{t_{i-1}+1} + X_i| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \lambda^{C' \log n} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}} \mid |\alpha_{t_{i-1}+1}| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} \right) \\ & \leq \mathbb{P} \left( |X_i| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \lambda^{C' \log n} \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}} \right). \end{aligned} \quad (47)$$

Here, the penultimate line follows from the independence relation stated above, whereas the last line follows from the elementary fact that

$$\mathbb{P}_{X \sim \mathcal{N}(\mu, \sigma^2)}(|X| < x) \leq \mathbb{P}_{X \sim \mathcal{N}(0, \sigma^2)}(|X| < x), \quad \forall x > 0.$$

Putting these pieces together, we conclude that

$$\begin{aligned} \mathbb{P} \left( |\alpha_k| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}, \text{ for all } k \leq 201C' \log n \right) & \leq \prod_{i=1}^{200} \mathbb{P} \left( |X_i| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \frac{\lambda^{C' \log n} \log^4 n}{n^{3/4}(\lambda-1)^{1.5}} \right) \\ & = \prod_{i=1}^{200} \mathbb{P} \left( \sqrt{\frac{n(\lambda^2-1)}{\lambda^{2C' \log n + 2} - \lambda^2}} |X_i| \lesssim \frac{(\lambda-1)n^{1/4}}{\lambda^{C' \log n}} + \frac{\log^4 n}{n^{1/4}(\lambda-1)} \right) \\ & \lesssim \left( \frac{\log^4 n}{n^{1/4}(\lambda-1)} \right)^{200} \lesssim n^{-11}, \end{aligned} \quad (48)$$

where we invoke the distribution of  $X_i$  in expression (45), and the last inequality results from the assumption that  $\lambda-1 \gtrsim n^{-1/9}$ . Therefore, the above inequality guarantees that with probability at least  $1 - O(n^{-10})$ , there exists some  $k \lesssim \frac{\log n}{\lambda-1}$  such that

$$|\alpha_k| \gtrsim \sqrt{\lambda-1} n^{-1/4}. \quad (49)$$

It thus implies that  $\tau_0 \lesssim \frac{\log n}{\lambda-1}$  (see the definition (41)), as claimed in (42). In other words, after at most  $O(\frac{\log n}{\lambda-1})$  iterations,  $|\alpha_t|$  shall surpass the order of  $\sqrt{\lambda-1} n^{-1/4}$ .  $\square$

**Stage II: moderate-to-large correlation** ( $\sqrt{\lambda-1} n^{-1/4} \lesssim |\alpha_t| \leq \frac{1}{2} \sqrt{\lambda^2-1}$ ). Next, let us look at the time interval after  $|\alpha_t|$  surpasses the level of  $\sqrt{\lambda-1} n^{-1/4}$  but before it reaches the level of  $\frac{1}{2} \sqrt{\lambda^2-1}$ . Mathematically, this refers to the interval  $(\tau_0, \varsigma]$ , where  $\tau_0$  and  $\varsigma$  are defined in (41) and (6), respectively. In fact, we shall start by examining

$$t \in \left( \tau_0, \varsigma \wedge \frac{c_5 \log n}{\lambda-1} \right] \quad (50)$$

for some constant  $c_5 > 0$ ; we shall demonstrate that  $\varsigma \lesssim \frac{\log n}{\lambda-1}$  shortly.

In view of Theorem 2 and the bounds (30), we can write

$$\alpha_{t+1} = \lambda v^{\star\top} \int \eta_t \left( \alpha_t v^{\star} + \frac{1}{\sqrt{n}} x \right) \varphi_n(dx) + \Delta_{\alpha,t}, \quad (51)$$

where the residual term obeys

$$|\Delta_{\alpha,t}| \lesssim \sqrt{\frac{t \log n}{n}} + |v^{\star\top} \eta_t(x_t) - v^{\star\top} \eta_t(v_t)|. \quad (52)$$

We first make a claim concerned with a refined recursive relation for  $\alpha_{t+1}$ :

$$|\alpha_{t+1}| \geq \frac{\lambda |\alpha_t|}{\sqrt{\alpha_t^2 + 1}} + o((\lambda - 1)|\alpha_t|) + O(|\Delta_{\alpha,t}|); \quad (53)$$

the proof of this result is postponed to Section A.5. Observe that whenever  $\alpha_t < \frac{1}{2}\sqrt{\lambda^2 - 1}$ , it holds that

$$\left( \frac{\lambda}{\sqrt{1 + \alpha_t^2}} \right)^2 \geq \frac{\lambda^2}{\frac{1}{4}\lambda^2 + \frac{3}{4}} > 1 + \frac{1}{3}(\lambda - 1), \quad \text{for } \lambda \in (1, 1.2], \quad (54)$$

which when taken together with expression (53), implies

$$|\alpha_{t+1}| \geq \left( \sqrt{1 + \frac{1}{3}(\lambda - 1)} + o(\lambda - 1) \right) |\alpha_t| + O(|\Delta_{\alpha,t}|). \quad (55)$$

In addition, we claim that for every  $t \leq \tau'$  where  $\tau' := \min\{t : \alpha_t \geq (\lambda - 1)^{-3/4}n^{-1/4}\}$ , it satisfies

$$|\Delta_{\alpha,t}| \ll (\lambda - 1)|\alpha_t|, \quad (56)$$

which we shall establish in Section A.6. With the relations (55) and (56) in place, it obeys  $|\alpha_{\tau'+1}| \geq |\alpha_{\tau'}|$ . Moreover, observe that the bound (52) taken together with (36) ensures that

$$|\Delta_{\alpha,\tau'+1}| \lesssim \sqrt{\frac{(\tau' + 1)^3 \log n}{n}} \ll (\lambda - 1)|\alpha_{\tau'+1}|. \quad (57)$$

Invoking this argument recursively, we thus arrive at

$$|\alpha_{t+1}| \geq \left( \sqrt{1 + \frac{1}{3}(\lambda - 1)} + o(\lambda - 1) \right) |\alpha_t|.$$

Now taking the above recursive relation collectively with the assumption  $\lambda - 1 \gtrsim n^{-1/9}$  reveals that  $|\alpha_t|$  surpasses  $\frac{1}{2}\sqrt{\lambda^2 - 1}$  within at most  $O\left(\frac{\log n}{\lambda-1}\right)$  iterations. Therefore, recalling our definition (6) of  $\varsigma$ , we can readily conclude that

$$\varsigma = O\left(\frac{\log n}{\lambda - 1}\right). \quad (58)$$

### 3.2.3 A more complete bound for $\xi_t$ (Proof of Claim (25))

We now move on to establish claim (25) for any  $t$  obeying  $\varsigma < t < \frac{cn(\lambda-1)^5}{\log^2 n}$  (recall from (58) that  $\varsigma = O\left(\frac{\log n}{\lambda-1}\right)$ ), again via an inductive argument. Along the way, we also need to demonstrate that

$$|\alpha_t| \geq \frac{1}{2}\sqrt{\lambda^2 - 1} \quad (59)$$

within this stage (namely, once  $|\alpha_t|$  exceeds  $\frac{1}{2}\sqrt{\lambda^2 - 1}$ , the signal strength will never fall below this level).

To begin with, the claim (25) for the base case  $t = \varsigma$  has already been validated in expression (36); the condition (59) also holds with high probability when  $t = \varsigma$ . Next, assuming that the claim (25) holds up till iteration  $t - 1$ , we would like to establish its validity for time  $t$ . Towards this end, inequality (33) together with Lemma 4 tells us that

$$\begin{aligned} \|\xi_t\|_2 &\leq \sqrt{\kappa_t + \sqrt{\frac{t \log^2 n}{n}} \|\xi_{t-1}\|_2} + O\left(\sqrt{\frac{t \log n}{n}} + \sqrt{(1+t \mathbb{1}_{t \leq \varsigma}) \log n} \|\xi_{t-1}\|_2^2 + \sqrt{\frac{t \log n}{n}} \|\xi_{t-1}\|_2\right) \\ &\leq \sqrt{1 - \frac{1}{15}(\lambda - 1) + \sqrt{\frac{t \log^2 n}{n}} \|\xi_{t-1}\|_2} + O\left(\sqrt{\frac{t \log n}{n}}\right) \\ &\quad + O\left(\sqrt{(1+t \mathbb{1}_{t \leq \varsigma}) \log n} \left(\sqrt{\frac{t \mathbb{1}_{t > \varsigma} \log n}{n(\lambda - 1)^2}} + \sqrt{\frac{\min\{t, \varsigma\}^3 \log n}{n}}\right) + \sqrt{\frac{t \log n}{n}}\right) \|\xi_{t-1}\|_2 \\ &\leq \left(1 - \frac{1}{15}(\lambda - 1) + O\left(\sqrt{\frac{(\frac{t}{(\lambda-1)^2} + \varsigma^3) \log^2 n}{n}}\right)\right) \|\xi_{t-1}\|_2 + O\left(\sqrt{\frac{t \log n}{n}}\right), \end{aligned} \quad (60)$$

where the second line comes from (32b) and the induction hypothesis (25) for  $t - 1$ . In addition, the validity of (59) for  $\alpha_{t+1}$  in the  $(t + 1)$ -th iteration can be justified as well, which we shall detail in Section A.7.

With the above recursive relation in mind, recognizing  $\sqrt{\frac{(t/(\lambda-1)^2 + \varsigma^3) \log^2 n}{n}} \leq 2c(\lambda - 1)$  for some constant  $c > 0$  small enough, we can readily derive

$$\begin{aligned} \|\xi_t\|_2 &\leq \left(1 - \frac{1}{20}(\lambda - 1)\right) \|\xi_{t-1}\|_2 + O\left(\sqrt{\frac{t \log n}{n}}\right) \\ &\leq \left(1 - \frac{1}{20}(\lambda - 1)\right)^{t-\varsigma} \|\xi_\varsigma\|_2 + \sum_{j=0}^{t-\varsigma-1} \left(1 - \frac{1}{20}(\lambda - 1)\right)^j O\left(\sqrt{\frac{(t-j) \log n}{n}}\right) \\ &\lesssim \sqrt{\frac{\varsigma^3 \log n}{n}} + \sqrt{\frac{t \log n}{n(\lambda - 1)^2}} \end{aligned} \quad (61)$$

for all  $\varsigma \leq t \leq \frac{cn(\lambda-1)^5}{\log^2 n}$ . Combining this with the bound (36) (for  $t \leq \varsigma$ ) immediately establishes Claim (25) for all  $t \leq \frac{cn(\lambda-1)^5}{\log^2 n}$ .

### 3.2.4 Analysis for approximate state evolution (Proof of Property (10))

Once the signal strength  $\alpha_t$  reaches the order of  $\sqrt{\lambda^2 - 1}$ , AMP enters the stage of local refinement. According to (25) and (24), for any  $t \leq \frac{cn(\lambda-1)^5}{\log^2 n}$ , the AMP iterate  $x_t$  admits the decomposition (8a) with the error term bounded by

$$\|\xi_t\|_2 \lesssim \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda - 1)^2}}. \quad (62)$$

In the meantime, to describe how  $\alpha_t$  evolves, we bound  $\Delta_{\alpha,t}$  based on the relation (29a) as follows:

$$|\Delta_{\alpha,t}| \lesssim B_t + \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda - 1)^2}}, \quad (63)$$

where the last inequality arises from (30). Combining this with relation (28) leads to

$$\alpha_{t+1} = \lambda v^{\star \top} \int \eta_t \left( \alpha_t v^\star + \frac{1}{\sqrt{n}} x \right) \varphi_n(dx) + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda - 1)^2}}\right).$$

Next, we shall characterize the distance between  $\alpha_{t+1}$  and its asymptotic counterpart to further understand the evolution of  $\alpha_{t+1}$ . More specifically, recall that the asymptotic state evolution is defined as

$$\alpha_{t+1}^* = \lambda \left[ \int \tanh(\alpha_t^* (\alpha_t^* + x)) \varphi(dx) \right]^{1/2}, \quad (64)$$

assuming we start from  $\alpha_\zeta^* = |\alpha_\zeta|$  for some  $\zeta = O(\frac{\log n}{\lambda-1})$ . We aim to control the difference between  $\alpha_{t+1}$  and  $\alpha_{t+1}^*$ . To simplify the presentation, we assume without loss of generality that  $\alpha_t > 0$ , and employ the notation

$$\tau_t := (\alpha_t^*)^2.$$

To begin with, the same analysis as in the proof of claim (53) (with different error bound (63) here) gives

$$\begin{aligned} \alpha_{t+1} &= \lambda \left[ \int \tanh(\alpha_t^2 + \alpha_t x) \varphi(dx) \right]^{1/2} + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^3 + \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^2}}\right) \\ &= \lambda \left[ \int \tanh(\alpha_t^2 + \alpha_t x) \varphi(dx) \right]^{1/2} + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^2}}\right). \end{aligned} \quad (65)$$

Here, the last line follows from inequality (20a) which indicates

$$\left(\frac{\pi_t}{\sqrt{n}}\right)^2 = \alpha_t^2 + O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right) = \alpha_t^2 + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^2}}\right).$$

It then follows from relations (65) and (64) that

$$\frac{\alpha_{t+1}^2 - \tau_{t+1}}{\tau_{t+1}} = \frac{\int [\tanh(\alpha_t^2 + \alpha_t x) - \tanh(\tau_t + \sqrt{\tau_t} x)] \varphi(dx)}{\int \tanh(\tau_t + \sqrt{\tau_t} x) \varphi(dx)} + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^3}}\right). \quad (66)$$

Here, we remind the readers that (see also (113) and Deshpande et al. (2017, Appendix B.2))

$$\int \tanh(\alpha^2 + \alpha x) \varphi(dx) = \int \tanh^2(\alpha^2 + \alpha x) \varphi(dx) \asymp \alpha^2, \quad \text{for } \alpha \in (0, \lambda],$$

where the last inequality results from relation (113). The recursive formula (66) quantifies how the difference between  $\alpha_t$  and  $\alpha_t^*$  changes over time, which plays a key role in our following analysis.

In order to better understand the above recursion, let us define — for every  $\tau \in [0, \lambda^2]$  — that

$$h(\tau) := \int \tanh(\tau + \sqrt{\tau} x) \varphi(dx).$$

Armed with this function, one can write

$$\alpha_{t+1}^{*2} = \lambda^2 h(\alpha_t^{*2}). \quad (67)$$

Also, direct calculations yield

$$h'(\tau) := \int \left(1 + \frac{x}{2\sqrt{\tau}}\right) (1 - \tanh^2(\tau + \sqrt{\tau} x)) \varphi(dx) \in (0, 1),$$

where its range follows from display (263) in Li and Wei (2022, Section D.3.3). We make note of a few direct consequences of the above results.

- Recognizing that  $h'(\tau) > 0$ , one has  $\alpha_{t+1}^* > \alpha_t^* \gtrsim \sqrt{\lambda^2 - 1}$  for  $t \geq \zeta$ .

- In view of display (264) in Li and Wei (2022, Section D.3.3), we have  $0 \leq \lambda^2 h'(\tau) \leq 1 - (\lambda - 1)$ . If we define  $\alpha^*$  to be the limiting point of (64) (as  $t \rightarrow \infty$ ), we can then see that

$$|\alpha_{t+1}^{*2} - \alpha^{*2}| \leq (1 - (\lambda - 1)) \cdot |\alpha_t^{*2} - \alpha^{*2}|, \quad \text{for } t \geq \varsigma. \quad (68)$$

In other words, the asymptotic state evolution parameter  $\alpha_t^{*2}$  converges exponentially to some fixed point  $\alpha^{*2}$ .

- In light of the above notation, we can also write

$$\begin{aligned} \frac{|\alpha_{t+1}^2 - \tau_{t+1}|}{\tau_{t+1}} &= \frac{|h(\alpha_t^2) - h(\tau_t)|}{h(\tau_t)} + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^3}}\right) \\ &= \frac{h'(\tau)}{h(\tau_t)/\tau_t} \cdot \frac{|\alpha_t^2 - \tau_t|}{\tau_t} + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^3}}\right) \end{aligned} \quad (69)$$

for some  $\tau$  satisfying  $\min\{\tau_t, \alpha_t^2\} \leq \tau \leq \max\{\tau_t, \alpha_t^2\}$ .

We first prove that  $\alpha_t^2 = (1 + o(1))\tau_t$ . By definition,  $\alpha_\varsigma^* = \alpha_\varsigma \gtrsim \sqrt{\lambda^2 - 1}$ , and hence this claim holds trivially for  $t = \varsigma$ . Next, assuming the validity of the inductive assumption  $\alpha_t^2 = (1 + o(1))\tau_t$ , we would like to prove it for the  $(t + 1)$ -th step. Towards this end, we first claim that there exists some universal constant  $c > 0$  small enough such that

$$\frac{h'(\tau)}{h(\tau_t)/\tau_t} \leq 1 - c(\lambda - 1), \quad (70)$$

whose proof is postponed to Section A.9. This further allows us to derive

$$\frac{|\alpha_t^2 - \tau_t|}{\tau_t} \lesssim \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^5}}, \quad (71)$$

which we shall demonstrate via an inductive argument. Given that (71) holds trivially when  $t = \varsigma$ , we intend to establish (71) for the  $(t + 1)$ -th iteration, assuming that it holds for all  $s \leq t$ . To do so, we combine (70) and (69) to show that

$$\begin{aligned} \frac{|\alpha_{t+1}^2 - \tau_{t+1}|}{\tau_{t+1}} &\leq (1 - c(\lambda - 1)) \cdot \frac{|\alpha_t^2 - \tau_t|}{\tau_t} + O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^3}}\right) \\ &= (1 - c(\lambda - 1))^{t+1-\varsigma} \cdot \frac{|\alpha_\varsigma^2 - \tau_\varsigma|}{\tau_\varsigma} + O\left(\sum_{k=\varsigma}^{t-1} (1 - c(\lambda - 1))^{t-k} \sqrt{\frac{(k + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^3}}\right) \\ &\leq \sum_{k=\varsigma}^{t-1} (1 - c(\lambda - 1))^{t-k} \cdot O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^3}}\right) \\ &\lesssim \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^5}}, \end{aligned} \quad (72)$$

where the second line is obtained by applying (71) recursively. Hence, it also leads to  $\alpha_{t+1}^2 = (1 + o(1))\tau_{t+1}$ , which concludes the inductive assumption for the  $(t + 1)$ -th iteration. Putting the above results together with expression (68) also gives

$$\begin{aligned} |\alpha_{t+1}^2 - \alpha^{*2}| &= (1 - (\lambda - 1))^{t-\varsigma} \cdot |\alpha_\varsigma^{*2} - \alpha^{*2}| + \alpha_{t+1}^{*2} O\left(\sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^5}}\right) \\ &\lesssim (1 - (\lambda - 1))^{t-\varsigma} + \sqrt{\frac{(t + \frac{\log^3 n}{\lambda-1}) \log n}{n(\lambda-1)^5}}. \end{aligned} \quad (73)$$

## 4 Discussions

In this paper, we have pinned down the finite-sample convergence behavior of AMP when initialized randomly, focusing on the prototypical  $\mathbb{Z}_2$  synchronization problem. This algorithm has been shown to enjoy fast global convergence, as it takes no more than  $O(\frac{\log n}{\lambda-1})$  iterations to arrive at a point whose risk is  $O(\sqrt{\frac{\log^4 n}{n(\lambda-1)^6}})$  close to Bayes-optimal. To the best of our knowledge, our theory offers the first rigorous evidence supporting the effectiveness of randomly initialized AMP in low-rank matrix estimation. While the present paper concentrates on a specific choice of denoising functions tailored to  $\mathbb{Z}_2$  synchronization, we expect our analysis framework to be generalizable to a broader family of separable and Lipschitz-continuous denoising functions.

Moving forward, there is no shortage of research directions worth exploring. One natural extension is concerned with other structural prior about  $v^*$ ; for instance, it would be interesting to see how randomly initialized AMP performs when  $v^*$  is known to satisfy general cone constraints (see e.g. Bandeira et al. (2019); Wei et al. (2019)). Another direction of interest is to go beyond the spiked Wigner model. A recent work along this line (Wu and Zhou, 2022) studied the role of random initialization for power iteration in the problem of tensor decomposition, which leverages upon the AMP-type analysis for analyzing tensor power methods. Can we further extend these to understand (randomly initialized) AMP towards solving more challenging problems like low-rank matrix completion and tensor completion? Moreover, note that the current theory characterizes the behavior of AMP up to  $O(\frac{n}{\text{poly}(\log n)})$  number of steps. While it is already useful in practice, it would be of interest — from a technical perspective — to understand whether AMP eventually converges to a fixed point as the iteration number further increases. Finally, while AMP serves as a versatile machinery for understanding various statistical procedures in high dimensions, there are several alternative analysis frameworks like the convex Gaussian minmax theorem (CGMT) (Celentano et al., 2020; Miolane and Montanari, 2021; Thrampoulidis et al., 2018) and the leave-one-out analysis (Abbe et al., 2020; Chen et al., 2020; El Karoui, 2018) that also prove effective and enjoy their own benefits. Is there any effective way to combine them so as to exploit all of their advantages at once? We leave these questions for future investigation.

## Acknowledgment

This work was partially supported by NSF grants DMS 2147546/2015447 and the NSF CAREER award DMS-2143215.

## A Proof of auxiliary lemmas and claims

### A.1 Proof of Lemma 2

**Proof of property (20a).** Recall that  $\pi_t := \sqrt{n(\|x_t\|_2^2 - 1)} \vee 1$ . To show property (20a), the first step is to calculate  $\|x_t\|_2$ . Notice that for independent Gaussian vectors  $\phi_k \sim \mathcal{N}(0, \frac{1}{n}I_n)$ , one has

$$v^{*\top} [\phi_1, \dots, \phi_{t-1}] \sim \mathcal{N}\left(0, \frac{1}{n}I_{t-1}\right),$$

given that  $\|v^*\|_2 = 1$ . Therefore, it is easily seen that

$$\begin{aligned} \left| \left\langle v^*, \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\rangle \right| &= \left| \left\langle v^{*\top} [\phi_1, \dots, \phi_{t-1}], [\beta_{t-1}^1, \dots, \beta_{t-1}^{t-1}] \right\rangle \right| \leq \left\| v^{*\top} [\phi_1, \dots, \phi_{t-1}] \right\|_2 \|\beta_{t-1}\|_2 \\ &\lesssim \sqrt{\frac{t \log n}{n}} \end{aligned} \tag{74}$$

with probability at least  $1 - O(n^{-11})$ . Combining inequalities (74) and (86), we reach

$$\|v_t\|_2^2 = \left\| \alpha_t v^* + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_2^2 = \|\alpha_t v^*\|_2^2 + \left\| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_2^2 + 2 \left\langle \alpha_t v^*, \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\rangle$$

$$= \alpha_t^2 + 1 + O\left(\sqrt{\frac{t \log n}{n}}\right) \asymp 1, \quad (75)$$

which in turn leads to

$$\|x_t\|_2^2 = (\|v_t\|_2 + O(\|\xi_{t-1}\|_2))^2 = \alpha_t^2 + 1 + O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right). \quad (76)$$

Based on the above properties, we can also derive that

$$\begin{aligned} \sqrt{n(\|x_t\|_2^2 - 1)} &= \sqrt{n\left(\alpha_t^2 + O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)\right)} \leq \sqrt{n}|\alpha_t| + O\left(\sqrt{n}\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)\right) \\ &= \sqrt{n}\left(|\alpha_t| + O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}\right)\right). \end{aligned} \quad (77)$$

where the first inequality based on the fact that  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$  for  $a, b \geq 0$ . In particular, if  $\alpha_t^2 \lesssim \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}$ , then the basic relation  $\sqrt{a+b} = \sqrt{a} + O(\sqrt{b})$  ( $0 \leq a \lesssim b$ ) enables us to replace “ $\leq$ ” in (78) with “ $=$ ” to obtain

$$\sqrt{n(\|x_t\|_2^2 - 1)} = \sqrt{n}\left(\alpha_t + O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}\right)\right). \quad (78)$$

In contrast, if  $\alpha_t^2 \gtrsim \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}$ , then the basic relation  $\sqrt{a+b} = \sqrt{a} + O(\frac{b}{\sqrt{a}})$  ( $0 \leq b \lesssim a$ ) yields

$$\begin{aligned} \sqrt{n(\|x_t\|_2^2 - 1)} &= \sqrt{n\left(\alpha_t^2 + O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)\right)} = \sqrt{n}|\alpha_t| \sqrt{1 + \frac{1}{\alpha_t^2} O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)} \\ &= \sqrt{n}\left(|\alpha_t| + \frac{1}{|\alpha_t|} O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)\right). \end{aligned} \quad (79)$$

The preceding two bounds taken collectively demonstrate that

$$\sqrt{n(\|x_t\|_2^2 - 1)} = \sqrt{n}|\alpha_t| + \sqrt{n}\left\{\frac{1}{|\alpha_t|} O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right) \wedge O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}\right)\right\}.$$

The above bound also leads to the desired form (20a) for  $\pi_t$  when  $\sqrt{n(\|x_t\|_2^2 - 1)} \geq 1$ . Consequently, it remains to examine the case where  $\sqrt{n(\|x_t\|_2^2 - 1)} < 1$ , which clearly can only happen if

$$c_{10}|\alpha_t| \leq \left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}$$

for some sufficiently small constant  $c_{10} > 0$ . But in this situation, we still have

$$\begin{aligned} \pi_t &= 1 \lesssim \sqrt{n}\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2} \asymp \sqrt{n}\left(|\alpha_t| + O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}\right)\right) \\ &\asymp \sqrt{n}|\alpha_t| + \sqrt{n}\left\{O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}\right) \wedge \frac{1}{|\alpha_t|} O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)\right\}, \end{aligned}$$

and hence the claimed bound is still valid.

**Proof of property (20b).** First recall the definition  $\gamma_t^{-2} := \|\tanh(\pi_t x_t)\|_2^2$ . Towards establishing property (20b), consider the following difference

$$\begin{aligned} & \left| \|\tanh(\pi_t x_t)\|_2^2 - \int \left\| \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \right\|_2^2 \varphi_n(dx) \right| \\ & \leq \left| \|\tanh(\pi_t x_t)\|_2^2 - \|\tanh(\pi_t v_t)\|_2^2 \right| + \left| \|\tanh(\pi_t v_t)\|_2^2 - \int \left\| \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \right\|_2^2 \varphi_n(dx) \right| \end{aligned}$$

where  $\varphi_n \sim \mathcal{N}(0, I_n)$ . To bound the right-hand side above, note that the Lipschitz property of  $\tanh$  gives

$$\|\tanh(\pi_t x_t) - \tanh(\pi_t v_t)\|_2 \leq \|\tanh(\pi_t(x_t - v_t))\|_2 \leq \pi_t \|\xi_{t-1}\|_2.$$

This in turn yields

$$\begin{aligned} \left| \|\tanh(\pi_t x_t)\|_2^2 - \|\tanh(\pi_t v_t)\|_2^2 \right| &= \left| \|\tanh(\pi_t x_t)\|_2 + \|\tanh(\pi_t v_t)\|_2 \right| \cdot \left| \|\tanh(\pi_t x_t)\|_2 - \|\tanh(\pi_t v_t)\|_2 \right| \\ &\leq (2 \|\tanh(\pi_t v_t)\|_2 + \pi_t \|\xi_{t-1}\|_2) \pi_t \|\xi_{t-1}\|_2 \\ &\lesssim (\pi_t \|v_t\|_2 + \pi_t \|\xi_{t-1}\|_2) \pi_t \|\xi_{t-1}\|_2 \asymp \pi_t^2 \|\xi_{t-1}\|_2, \end{aligned}$$

where the last line makes use of the assumption  $\|\xi_{t-1}\|_2 \lesssim 1$  and the equation (75). Thus, we arrive at

$$\begin{aligned} & \left| \|\tanh(\pi_t x_t)\|_2^2 - \int \left\| \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \right\|_2^2 \varphi_n(dx) \right| \\ & \lesssim \pi_t^2 \|\xi_{t-1}\|_2 + \left| \|\tanh(\pi_t v_t)\|_2^2 - \int \left\| \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \right\|_2^2 \varphi_n(dx) \right|. \end{aligned} \quad (80)$$

Next we develop a bound on the second term of (80), which turns out to be a consequence of the uniform concentration property. Specifically, let us consider functions of the following form

$$f_\theta(\Phi) := \|\tanh(\pi v)\|_2^2 - \int \left\| \tanh \left( \frac{\pi}{\sqrt{n}} (\alpha + x) \right) \right\|_2^2 \varphi_n(dx), \quad \text{where } v = \alpha v^* + \sum_{k=1}^{t-1} \beta^k \phi_k;$$

here, we define

$$\Phi = \sqrt{n} [\phi_1, \phi_2, \dots, \phi_{t-1}], \quad \beta = [\beta^1, \beta^2, \dots, \beta^{t-1}], \quad \theta = [\alpha, \beta, \pi] \in \mathbb{R}^{t+1}.$$

Then, it suffices to bound  $f_\theta(\Phi)$  uniformly over all  $\theta$  in the following set

$$\Theta := \left\{ \theta = (\alpha, \beta, \pi) \mid \|\beta\|_2 = 1, \alpha \lesssim \sqrt{\lambda^2 - 1}, \pi \lesssim \sqrt{n} \right\}. \quad (81)$$

In order to achieve this, first consider the derivative of  $f$  with respect to  $\Phi$ , which by direct calculations satisfies

$$\begin{aligned} \|\nabla_\Phi f_\theta(\Phi)\|_2 &\leq \frac{2\pi \|\beta\|_2}{\sqrt{n}} \|\tanh(\pi v) \circ \tanh'(\pi v)\|_2 \\ &\leq \frac{2\pi \|\beta\|_2}{\sqrt{n}} \|\tanh(\pi v)\|_2 \lesssim \frac{\pi^2}{\sqrt{n}} \|v\|_2 \lesssim \frac{\pi^2}{\sqrt{n}}, \end{aligned} \quad (82)$$

where we note that  $\|v\|_2 \leq \alpha + \frac{1}{\sqrt{n}} \|\Phi\| \lesssim 1$ . Additionally, since function  $f_\theta(\Phi)$  is uniformly bounded by 2, if we take  $\delta = n^{-300}$  for the set  $\mathcal{E}$  defined in Lemma 1, it satisfies (we refer the readers to Li and Wei (2022, Section D.1.1) for the proof of this property)

$$\mathbb{E} [|f(\Phi) - f(\mathcal{P}_\mathcal{E}(\Phi))|] \lesssim n^{-100}.$$

where  $\mathcal{P}_\mathcal{E}(\cdot)$  denotes the Euclidean projection onto the set  $\mathcal{E}$ . Combining the above inequality with the following properties of function  $f_\theta(\Phi)$ ,

1.  $\|\nabla_\theta f_\theta(\Phi)\|_2 \lesssim n^{100}$  for any  $\Phi \in \mathcal{E}$ ,
2. For any fixed  $\theta$ , one has  $\mathbb{E}[f_\theta(\Phi)] = 0$ ,

we can apply [Li and Wei \(2022, Corollary 3\)](#) to reach

$$\sup_{\theta \in \Theta} \left| \frac{1}{\pi^2} f_\theta(\Phi) \right| \lesssim \sqrt{\frac{t \log n}{n}}. \quad (83)$$

Taking everything collectively, we conclude that

$$\left| \|\tanh(\pi_t x_t)\|_2^2 - \int \left\| \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \right\|_2^2 \varphi_n(dx) \right| \lesssim \pi_t^2 \left( \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \right), \quad (84)$$

which leads to the property (20b) by recognizing that

$$\int \left\| \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \right\|_2^2 \varphi_n(dx) = n \int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx).$$

**Proof of property (20c).** We are only left with calculating the value of  $\int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx)$  which shall be done as follows. First, by observing that  $\tanh(0) = \tanh''(0) = 0$ ,  $\tanh'(0) = 1$ ,  $|\tanh'''(x)| \leq 4$  for  $x \in \mathbb{R}$ , we find

$$|\tanh(x) - x| \leq \frac{2}{3}|x|^3,$$

as a consequence of the mean value theorem. Combining this relation with the fact that  $|\tanh(x) - x| \leq |x|$ , we can further obtain

$$|\tanh(x) - x| \leq |x| \wedge |x|^3.$$

As a result, we see that

$$|\tanh(x) + x| = |2x + \tanh(x) - x| = 2|x| + O(|x| \wedge |x|^3) \lesssim |x|,$$

which in turn leads to

$$\tanh^2(x) = x^2 + O(|\tanh(x) - x| |\tanh(x) + x|) = x^2 + O(x^4).$$

Now we are ready to compute the value of  $\int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx)$ . In view of the expressions obtained above, it follows that

$$\begin{aligned} \int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx) &= \int \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right)^2 \varphi(dx) + O \left( \int \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right)^4 \varphi(dx) \right) \\ &= \frac{\pi_t^2}{n} (\alpha_t^2 + 1) + O \left( \frac{\pi_t^4}{n^2} \right). \end{aligned} \quad (85)$$

We thus complete the proof of the advertised result.

## A.2 Proof of Lemma 3

**Proof of property (22a).** The property (22a) is concerned with the magnitudes of  $\eta_t$  and its derivatives. In view of the definition of  $\eta_t$ , we proceed to bounding the parameters  $\pi_t$  and  $\gamma_t$  separately. Towards this, recall from [Lemma 1](#) that: with probability at least  $1 - O(n^{-11})$ ,

$$\left\| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_2 = 1 + O \left( \sqrt{\frac{t \log n}{n}} \right) \quad (86)$$

holds simultaneously for all  $\beta_{t-1} = [\beta_{t-1}^k]_{1 \leq k < t} \in \mathcal{S}^{t-2}$ . It then follows from this result and the definition (19) of  $v_t$  that

$$\|v_t\|_2 = \left\| \alpha_t v^\star + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_2 \leq \|\alpha_t v^\star\|_2 + \left\| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_2 = |\alpha_t| + 1 + O\left(\sqrt{\frac{t \log n}{n}}\right) \lesssim 1, \quad (87)$$

given that  $t \lesssim \frac{n}{\log n}$  and

$$|\alpha_t| = |\lambda v^\star \top \eta_{t-1}(x_{t-1})| \leq \lambda \lesssim 1. \quad (88)$$

These properties together with the assumption on  $\xi_t$  enable us to control the  $\ell_2$  norm of  $x_t$  as follows:

$$\|x_t\|_2 = \|v_t + \xi_{t-1}\|_2 \leq \|v_t\|_2 + \|\xi_{t-1}\|_2 \lesssim 1. \quad (89)$$

Given that  $\|x_t\|_2 \lesssim 1$ , the value of  $\pi_t$  can be controlled as

$$\pi_t := \sqrt{n(\|x_t\|_2^2 - 1)} \vee 1 \lesssim \sqrt{n}. \quad (90)$$

Additionally, by observing that  $\tanh(0) = 0$ ,  $\tanh'(x) = 1 - \tanh^2(x) \in [0, 1]$  and  $|\tanh(x)| \leq 1$ , one has

$$|\tanh(x)| \asymp |x| \wedge 1. \quad (91)$$

We claim that this leads to the following consequence:

$$\|\tanh(\pi_t x_t)\|_2 \asymp \pi_t; \quad (92)$$

for the moment, let us first take this as given and we shall come back to its proof after establishing the property (22a). In view of this claim (92), we find that

$$\gamma_t := \|\tanh(\pi_t x_t)\|_2^{-1} \asymp \pi_t^{-1}. \quad (93)$$

In order to prove property (22a), it suffices to recall the definition of  $\eta_t$  as in expression (3). For any  $x \in \mathbb{R}$ , direct calculations give

$$\eta_t(x) = \gamma_t \tanh(\pi_t x) \quad (94a)$$

$$\eta'_t(x) = \gamma_t \pi_t (1 - \tanh^2(\pi_t x)) \quad (94b)$$

$$\eta''_t(x) = -2\gamma_t \pi_t^2 \tanh(\pi_t x) (1 - \tanh^2(\pi_t x)) \quad (94c)$$

$$\eta'''_t(x) = -2\gamma_t \pi_t^3 (1 - \tanh^2(\pi_t x)) (1 - 3\tanh^2(\pi_t x)). \quad (94d)$$

Combining these identities with (91), (93) and the fact that  $|\tanh(x)| \leq 1$ , one can easily validate expression (22a). It then boils down to justifying the claim (92), which we accomplish below.

**Proof of relation (92).** Note that from display (91), one has  $\|\tanh(\pi_t x_t)\|_2 \asymp \|\pi_t x_t \wedge 1\|_2$ , where both operators  $|\cdot|$  and  $\wedge$  are applied in an entrywise manner and we overlaid the notation 1 to denote an all-one vector. To establish the relation (92), it is sufficient to prove that  $\|\pi_t x_t \wedge 1\|_2 \asymp \pi_t$ . Towards this end, first we invoke (90) to make the observation that

$$\|\pi_t x_t \wedge 1\|_2 \leq \|\pi_t x_t\|_2 \wedge \|1\|_2 = \|\pi_t x_t\|_2 \wedge \sqrt{n} \asymp \pi_t \wedge \sqrt{n} \asymp \pi_t.$$

In addition, let us introduce an index set  $\mathcal{I}$  as follows:

$$\mathcal{I} := \left\{ i \in [n] \mid |\xi_{t-1,i}| \leq 0.9|v_{t,i}| \text{ and } |v_{t,i}| \lesssim \frac{1}{\sqrt{n}} \right\}, \quad (95)$$

which clearly satisfies

$$0.1|\pi_t v_{t,i}| \leq |\pi_t x_{t,i}| \leq 1.9|\pi_t v_{t,i}|, \quad \forall i \in \mathcal{I}.$$

It then follows that:

$$\|\pi_t x_t \wedge 1\|_2 \geq \|\pi_t x_t \circ \mathbf{1}_{\mathcal{I}} \wedge 1\|_2 \asymp \|\pi_t v_t \circ \mathbf{1}_{\mathcal{I}} \wedge 1\|_2. \quad (96)$$

To further control the right-hand side of display (96), we claim that

$$\|\pi_t v_t \circ \mathbf{1}_{\mathcal{I}} \wedge 1\|_2 \stackrel{(i)}{\asymp} \|\pi_t v_t \circ \mathbf{1}_{\mathcal{I}}\|_2 \stackrel{(ii)}{\asymp} \pi_t. \quad (97)$$

In order to see this, relation (i) can be verified using expression (90) and the definition of (95), as they guarantee that

$$|\pi_t v_{t,i}| \leq |\pi_t| \cdot |v_{t,i}| \lesssim 1, \quad \forall i \in \mathcal{I}.$$

To validate (ii), note that on the index set  $\mathcal{I}^c$ , one has  $|v_{t,i}| \lesssim |\xi_{t-1,i}|$ . Therefore, it holds that

$$\|\pi_t v_t \circ \mathbf{1}_{\mathcal{I}^c}\|_2 \lesssim \|\pi_t \xi_{t-1} \circ \mathbf{1}_{\mathcal{I}^c}\|_2 \lesssim \pi_t \|\xi_{t-1}\|_2 \lesssim \pi_t \sqrt{\frac{1}{\log n}}, \quad (98)$$

where we recall our assumption  $\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{1}{\log n}}$ . Equipped with the above calculation, we can recall  $\|v_t\|_2 \asymp 1$  from expression (87) to obtain

$$\|\pi_t v_t \circ \mathbf{1}_{\mathcal{I}}\|_2^2 = \|\pi_t v_t\|_2^2 - \|\pi_t v_t \circ \mathbf{1}_{\mathcal{I}^c}\|_2^2 \asymp \left(1 - O\left(\frac{1}{\log n}\right)\right) \pi_t^2 \asymp \pi_t^2, \quad (99)$$

as claimed in Part (ii) of (97).

**Proof of property (22b).** To study the derivatives of  $\eta_t(x_t)$ , we first consider the parameters  $\pi_t$  and  $\gamma_t$ . Given that  $\|\xi_{t-1}\|_2$  satisfies the expression (21), when  $t \lesssim \frac{\log n}{\lambda-1}$ , it holds true that

$$\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}} \lesssim \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}. \quad (100)$$

Under the assumption  $\lambda-1 \gtrsim n^{-1/9} \log n$ , we see that  $\|\xi_{t-1}\|_2$  also satisfies

$$\|\xi_{t-1}\|_2 \lesssim (\lambda-1) \cdot \sqrt{\frac{\log^4 n}{n} \cdot \frac{1}{(\lambda-1)^5}} \lesssim (\lambda-1) \cdot \sqrt{\frac{\log^4 n}{n} \cdot \frac{n^{5/9}}{\log^5 n}} \lesssim (\lambda-1)n^{-0.2}.$$

Similarly, it is also easily seen that

$$\left(\frac{t \log n}{n}\right)^{1/2} \lesssim (\lambda-1)n^{-0.2}.$$

Taking these together with the relation (78) in the proof of Lemma 2 and the assumption  $\alpha_t \lesssim \sqrt{\lambda-1} n^{-0.1}$  yields

$$\pi_t \leq \sqrt{n} \alpha_t + O\left(\sqrt{n} \left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}\right) \lesssim \sqrt{\lambda-1} n^{0.4}. \quad (101)$$

Similarly, the relation (20d) of Lemma 2 combined with the assumption  $\alpha_t \leq \sqrt{\lambda-1} n^{-0.1}$  and the above bounds leads to

$$(\gamma_t \pi_t)^{-2} = \alpha_t^2 + 1 + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right) = 1 + O((\lambda-1)n^{-0.2}),$$

which by direct calculation also gives  $\gamma_t \pi_t = 1 + O((\lambda-1)n^{-0.2})$  under our assumption on  $\lambda-1$ .

Armed with the above properties, some algebra together with (94) further results in

$$\begin{aligned} \eta'_t(x) &= \gamma_t \pi_t (1 - \tanh^2(\pi_t x)) = 1 + O((\lambda-1)n^{-0.2} \log n), \\ |\eta''_t(x)| &\lesssim \pi_t \cdot |\pi_t x| \lesssim (\lambda-1)n^{0.8}|x|, \end{aligned}$$

where we invoke the relation  $|1 - \tanh^2(\pi_t x)| \asymp 1$  for any  $|x| \lesssim \sqrt{\frac{\log n}{n}}$ .

**Proof of property (22c).** Again when  $t \lesssim \frac{\log n}{\lambda-1}$ ,  $\|\xi_{t-1}\|_2$  satisfies inequality (100). Taking this collectively with the assumption  $\alpha_t \lesssim \sqrt{\lambda-1}n^{-1/4}$  and the relation (78) in the proof of Lemma 2, we obtain

$$\begin{aligned}\pi_t &\leq \sqrt{n}\alpha_t + O\left(\sqrt{n}\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}\right) \lesssim \sqrt{\lambda-1}n^{1/4} + (\lambda-1)^{-3/4}n^{1/4}\log n \\ &\asymp (\lambda-1)^{-3/4}n^{1/4}\log n.\end{aligned}\tag{102}$$

Similarly, the relation (20d) of Lemma 2 together with our assumption on  $\lambda-1$  yields

$$\gamma_t\pi_t = 1 + O\left(\frac{\log^2 n}{\sqrt{n}(\lambda-1)^3}\right).\tag{103}$$

To derive property (22c), we make note of some simple facts that  $\tanh(0) = \tanh''(0) = \tanh'''(0) = 0$ ,  $\tanh'(0) = 1$ ,  $\tanh'''(0) = -2$  and  $|\tanh^{(5)}(x)| \leq K$  for some constant  $K$ . As a result, the mean value theorem ensures that for any  $x$ , there exists a quantity  $c$  such that

$$\tanh(x) = x - \frac{1}{3}x^3 + cx^5 \quad \text{for some } 0 \leq c \leq K',$$

for  $K' = K/120$ . Based on the calculations above, we can conclude that

$$\eta_t(x) = \gamma_t \tanh(\pi_t x) = (1 - c_0)\pi_t^{-1} \tanh(\pi_t x) = (1 - c_0)\left(x - \frac{1}{3}\pi_t^2 x^3 + c_x\right),$$

where  $c_0$  and  $c_x$  are some quantities obeying

$$|c_0| \lesssim \frac{\log^2 n}{\sqrt{n}(\lambda-1)^3} \quad \text{and} \quad |c_x| \lesssim \pi_t^4 |x|^5 \lesssim \frac{n|x|^5 \log^4 n}{(\lambda-1)^3}.$$

This completes the proof of the desired property.

### A.3 Proof of Lemma 4

Without loss of generality, throughout this proof, we assume  $\alpha_t > 0$ . Before we begin to bound  $\kappa_t$ , let us simplify the term of interest a little bit. First of all, as each entry follows  $v_i^\star \stackrel{\text{i.i.d.}}{\sim} \text{Unif}\left\{\frac{1}{\sqrt{n}}, -\frac{1}{\sqrt{n}}\right\}$ , one can easily derive

$$\begin{aligned}&\left\langle \int \left[ x\eta'_t \left( \alpha_t v^\star + \frac{1}{\sqrt{n}}x \right) - \frac{1}{\sqrt{n}}\eta''_t \left( \alpha_t v^\star + \frac{1}{\sqrt{n}}x \right) \right]^2 \varphi_n(dx) \right\rangle \\ &= \int \left[ x\eta'_t \left( \alpha_t v_1^\star + \frac{1}{\sqrt{n}}x \right) - \frac{1}{\sqrt{n}}\eta''_t \left( \alpha_t v_1^\star + \frac{1}{\sqrt{n}}x \right) \right]^2 \varphi(dx) \\ &= \int \left[ x\eta'_t \left( \frac{1}{\sqrt{n}}(\alpha_t + x) \right) - \frac{1}{\sqrt{n}}\eta''_t \left( \frac{1}{\sqrt{n}}(\alpha_t + x) \right) \right]^2 \varphi(dx),\end{aligned}$$

where  $\varphi(\cdot)$  is the p.d.f. of  $\mathcal{N}(0, 1)$ , and we have used the fact that  $\eta_t(\cdot)$  is symmetric about 0 and the integration is over the distribution  $\mathcal{N}(0, I_n)$ . Similarly, we obtain

$$\left\langle \int \left[ \eta'_t \left( \alpha_t v^\star + \frac{1}{\sqrt{n}}x \right) \right]^2 \varphi_n(dx) \right\rangle = \int \left[ \eta'_t \left( \frac{1}{\sqrt{n}}(\alpha_t + x) \right) \right]^2 \varphi(dx).$$

Therefore, it holds that

$$\kappa_t^2 = \max \left\{ \int |I_1(x)|^2 \varphi(dx), \int |I_2(x)|^2 \varphi(dx) \right\},\tag{104}$$

where we define

$$\begin{aligned} I_1(x) &:= x\eta'_t\left(\frac{1}{\sqrt{n}}(\alpha_t + x)\right) - \frac{1}{\sqrt{n}}\eta''_t\left(\frac{1}{\sqrt{n}}(\alpha_t + x)\right), \\ I_2(x) &:= \eta'_t\left(\frac{1}{\sqrt{n}}(\alpha_t + x)\right). \end{aligned}$$

We now proceed to the proof of Lemma 4, and begin by restricting our attention to the range  $t < \varsigma \wedge \frac{\log n}{c(\lambda-1)}$ . We divide into two cases depending on the value of  $\alpha_t$ .

**Case I:**  $\alpha_t \lesssim \sqrt{\lambda^2 - 1} n^{-0.1}$ . Let us introduce an event  $\mathcal{A} := \{x : |x| \leq \sqrt{24 \log n}\}$ . For  $x \sim \mathcal{N}(0, 1)$ , it is easily verified that  $P(\mathcal{A}) = 1 - O(n^{-12})$ . Conditional on  $\mathcal{A}$  and assuming  $\alpha_t \lesssim \sqrt{\lambda^2 - 1} n^{-0.1}$ , one has

$$\left| \frac{1}{\sqrt{n}}(\alpha_t + x) \right| \lesssim \sqrt{\frac{\log n}{n}}.$$

Meanwhile, recall that  $\sqrt{\lambda+1} \asymp 1$ , and therefore  $\alpha_t \lesssim \sqrt{\lambda^2 - 1} n^{-0.1}$  is equivalent to  $\alpha_t \lesssim \sqrt{\lambda-1} n^{-0.1}$ . As a result, according to the property (22b) established in Lemma 3, we see that: when  $|z| \lesssim \sqrt{\frac{\log n}{n}}$ , one has

$$\eta'_t(z) = 1 + O((\lambda - 1)n^{-0.2} \log n) \quad \text{and} \quad |\eta''_t(z)| \lesssim (\lambda - 1)n^{0.8}|z|. \quad (105)$$

Hence, for all  $x$  residing within  $\mathcal{A}$ , we can bound the difference between  $I_1(x)$  and  $x$  uniformly as follows:

$$\begin{aligned} |I_1(x) - x| &\leq \left| \eta'_t\left(\frac{1}{\sqrt{n}}(\alpha_t + x)\right) - 1 \right| |x| + \frac{1}{\sqrt{n}} \left| \eta''_t\left(\frac{1}{\sqrt{n}}(\alpha_t + x)\right) \right| \\ &\lesssim ((\lambda - 1)n^{-0.2} \log n) |x| + (\lambda - 1)n^{0.3} \left| \frac{1}{\sqrt{n}}(\alpha_t + x) \right| \\ &\lesssim ((\lambda - 1)n^{-0.2} \log n) |x| + (\lambda - 1)n^{0.3} \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}}|x| \right) \\ &\lesssim ((\lambda - 1)n^{-0.2} \log n) (1 + |x|). \end{aligned} \quad (106)$$

In addition, consider any  $x \in \mathbb{R}$ . Recalling the relation (22a) of Lemma 3 — which reveals that  $|\eta'_t(x)| \lesssim 1$ ,  $|\eta''_t(x)| \lesssim \sqrt{n}$  — we observe the crude bound that  $|I_1(x)| \lesssim |x| + 1$ . Putting the preceding two bounds together, we arrive at

$$\begin{aligned} \int |I_1(x)|^2 \varphi(dx) &= \int x^2 \varphi(dx) + \int (I_1(x) - x)(I_1(x) + x) \varphi(dx) \\ &= 1 + \int_{\mathcal{A}} (I_1(x) - x)(I_1(x) + x) \varphi(dx) + \int_{\mathcal{A}^c} (I_1(x) - x)(I_1(x) + x) \varphi(dx) \\ &= 1 + O\left(\int_{\mathcal{A}} (\lambda - 1)n^{-0.2} \log n \cdot (|x| + 1)^2 \varphi(dx) + \int_{\mathcal{A}^c} (|x| + 1)^2 \varphi(dx)\right) \\ &= 1 + O((\lambda - 1)n^{-0.2} \log n), \end{aligned} \quad (107)$$

where the last equality utilizes the fact that for  $c_n = \sqrt{24 \log n}$ , one has

$$\int_{\mathcal{A}^c} x^2 \varphi(dx) = 2 \int_{c_n}^{\infty} x^2 \varphi(dx) \lesssim \int_{c_n}^{\infty} x^2 \exp\left(-\frac{1}{2}x^2\right) dx \lesssim \frac{\log n}{n^{12}}.$$

Regarding the other term  $I_2(x)$ , relation (22a) of Lemma 3 implies that  $|I_2(x)| \lesssim 1$ . Furthermore, if  $|x| \leq \sqrt{24 \log n}$ , in view of the relation (22b) we have  $|I_2(x)| = 1 + O((\lambda - 1)n^{-0.2} \log n)$ . Putting these together, we obtain

$$\begin{aligned} \int |I_2(x)|^2 \varphi(dx) &= \int_{\mathcal{A}} |I_2(x)|^2 \varphi(dx) + \int_{\mathcal{A}^c} |I_2(x)|^2 \varphi(dx) \\ &= 1 + O((\lambda - 1)n^{-0.2} \log n) + O(P(\mathcal{A}^c)) \\ &= 1 + O((\lambda - 1)n^{-0.2} \log n). \end{aligned} \quad (108)$$

Combining inequalities (107) and (108) then leads to

$$\kappa_t^2 = 1 + O((\lambda - 1)n^{-0.2} \log n) = 1 + o\left(\frac{\lambda - 1}{\log n}\right).$$

**Case II:**  $\sqrt{\lambda^2 - 1}n^{-0.1} \lesssim \alpha_t \lesssim \sqrt{\lambda^2 - 1}$ . We first note that under the assumption (23), the following relation holds for  $\|\xi_{t-1}\|_2$  when  $t \lesssim \frac{\log n}{\lambda - 1}$ :

$$\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}} \lesssim \sqrt{\frac{\log^4 n}{n(\lambda - 1)^3}}. \quad (109)$$

We recall the basic facts obtained in property (94), and as a result,

$$\int |I_1(x)|^2 \varphi(dx) = \int \left[ \left( \gamma_t \pi_t x + \frac{2}{\sqrt{n}} \gamma_t \pi_t^2 \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \right) \cdot \left( 1 - \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \right) \right]^2 \varphi(dx) \quad (110)$$

$$\int |I_2(x)|^2 \varphi(dx) = \int \left[ \gamma_t \pi_t (1 - \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right)) \right]^2 \varphi(dx). \quad (111)$$

It then comes down to controlling the right-hand side of the above two expressions, under the condition that  $\sqrt{\lambda^2 - 1}n^{-0.1} \lesssim \alpha_t \leq \sqrt{\lambda^2 - 1}$ .

For notational convenience, let us introduce additional shorthand notation as follows:

$$\begin{aligned} J_1(x) &:= \gamma_t \pi_t x + \frac{2}{\sqrt{n}} \gamma_t \pi_t^2 \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right), \\ J_2(x) &:= \mu_t \alpha_t x + 2\mu_t \alpha_t^2 \tanh\left(\alpha_t(\alpha_t + x)\right), \\ K_1(x) &:= 1 - \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right), \\ K_2(x) &:= 1 - \tanh^2\left(\alpha_t(\alpha_t + x)\right) \\ \mu_t &:= \left[ \int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx) \right]^{-\frac{1}{2}}. \end{aligned}$$

With this set of notation in place, it follows from (110) and a little algebra that

$$\begin{aligned} \int |I_1(x)|^2 \varphi(dx) &= \int [J_1(x)K_1(x)]^2 \varphi(dx) \\ &= \int [J_2(x)K_2(x)]^2 \varphi(dx) + O\left(\int |J_2(x)|^2 |K_1(x) - K_2(x)| |K_1(x) + K_2(x)| \varphi(dx)\right) \\ &\quad + O\left(\int |J_1(x) - J_2(x)| |J_1(x) + J_2(x)| |K_1(x)|^2 \varphi(dx)\right). \end{aligned} \quad (112)$$

To bound  $\int |I_1(x)|^2 \varphi(dx)$ , it is thus sufficient to control these three terms on the right-hand side of (112) separately. Before proceeding, we find it helpful to make note of several preliminary properties.

- First, we would like to show that

$$\mu_t \asymp \alpha_t^{-1}. \quad (113)$$

In order to see this, note that the elementary fact  $\tanh^2(x) \leq x^2$  implies that

$$\int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx) \leq \int \alpha_t^2 (\alpha_t + x)^2 \varphi(dx) \asymp \alpha_t^2, \quad \text{for } \alpha_t \leq \lambda.$$

On the other hand, when  $|x| \leq 1/2$ ,  $\alpha_t \in (0, \lambda]$  and  $\lambda \in (1, 1.2]$ , one has  $\alpha_t(\alpha_t + x) \in [-0.0625, 2.04]$ . Clearly, for any  $z \in [-0.0625, 2.04]$ , we have  $\tanh^2(z)/z^2 \geq 0.22$ , and as a consequence,

$$\begin{aligned} \int \tanh^2(\alpha_t(\alpha_t + x))\varphi(dx) &\geq \int \tanh^2(\alpha_t(\alpha_t + x))\mathbf{1}(|x| \leq 0.5)\varphi(dx) \\ &\gtrsim \int \alpha_t^2(\alpha_t + x)^2 \mathbf{1}(|x| \leq 0.5)\varphi(dx) \asymp \alpha_t^2. \end{aligned}$$

The preceding two bounds taken collectively justify the claim (113).

- Additionally, based on equation (20a) of Lemma 2 and the bound (109), we have

$$\pi_t = \sqrt{n}\alpha_t + \frac{\sqrt{n}}{\alpha_t}O\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right) = \alpha_t\sqrt{n} + O\left(\sqrt{\frac{\log^4 n}{\alpha_t^2(\lambda-1)^3}}\right) \quad (114)$$

$$= \alpha_t\sqrt{n} \left(1 + O\left(\sqrt{\frac{\log^4 n}{n\alpha_t^4(\lambda-1)^3}}\right)\right) \asymp \alpha_t\sqrt{n} \quad (115)$$

as long as  $t \lesssim \frac{\log n}{\lambda-1}$ . Here, the last inequality holds since

$$\sqrt{\frac{\log^4 n}{(\lambda-1)^3}} \lesssim (\lambda-1) \cdot \frac{\log^2 n}{(\lambda-1)^{5/2}} = o((\lambda-1)n^{0.3}) = o(\alpha_t^2\sqrt{n}),$$

provided that  $\alpha_t \gtrsim \sqrt{\lambda^2 - 1}n^{-0.1}$  and  $\lambda-1 \gtrsim n^{-1/9} \log n$ .

- Moreover, by combining equation (20b) of Lemma 2 with (114), (115) and (109), we see that

$$\begin{aligned} \gamma_t^{-2} &= n \int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(dx) + O\left(\alpha_t^2 \sqrt{\frac{n \log^4 n}{(\lambda-1)^3}}\right) \\ &= n \int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx) + O\left(\sqrt{\frac{n \log^4 n}{(\lambda-1)^3}}\right) \\ &= n\mu_t^{-2} + O\left(\sqrt{\frac{n \log^4 n}{(\lambda-1)^3}}\right). \end{aligned}$$

Here, the second equality arises from the following fact (see also (146)):

$$\begin{aligned} &\left| \int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx) \right| \\ &\leq \left|1 - \frac{\alpha_t^2 n}{\pi_t^2}\right| \int \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \varphi(dx) + \left| \int \frac{\alpha_t^2 n}{\pi_t^2} \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx) \right| \\ &\lesssim \left|1 - \frac{\alpha_t^2 n}{\pi_t^2}\right| \alpha_t^2 + \left| \frac{\pi_t^2}{\alpha_t^2 n} - 1 \right| \alpha_t^4 = O\left(\sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right). \end{aligned} \quad (116)$$

As a result, we can express  $\gamma_t$  in term of  $\mu_t$  as

$$\gamma_t = \left[n\mu_t^{-2} + O\left(\sqrt{\frac{n \log^4 n}{(\lambda-1)^3}}\right)\right]^{-1/2} = \frac{\mu_t}{\sqrt{n}} \left(1 + O\left(\mu_t^2 \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right)\right) \asymp \frac{\mu_t}{\sqrt{n}}. \quad (117)$$

- With (115) and (117) in place, a little algebra further leads to

$$|\gamma_t \pi_t - \mu_t \alpha_t| \lesssim \mu_t \alpha_t \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \left( \frac{1}{\alpha_t^2} + \mu_t^2 \right) \lesssim \frac{1}{\alpha_t^2} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}, \quad (118)$$

$$\left| \frac{1}{\sqrt{n}} \gamma_t \pi_t^2 - \mu_t \alpha_t^2 \right| \lesssim \mu_t \alpha_t^2 \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \left( \frac{1}{\alpha_t^2} + \mu_t^2 \right) \lesssim \frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}. \quad (119)$$

In addition, according to the relation (20d) in Lemma 2, we have

$$\gamma_t^2 \pi_t^2 = (\alpha_t^2 + 1)^{-1} + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right),$$

and using similar analysis as for (20c) yields

$$\mu_t^{-2} = \int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx) = \alpha_t^2(\alpha_t^2 + 1) + O(\alpha_t^4).$$

These two bounds taken together with a little algebra leads to

$$|\gamma_t \pi_t - \mu_t \alpha_t| \lesssim \frac{\pi_t^2}{n} + \alpha_t^2 + \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \lesssim \alpha_t^2 + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \lesssim \alpha_t^2, \quad (120)$$

recognizing the range  $\sqrt{\lambda^2 - 1} n^{-0.1} \lesssim \alpha_t \lesssim \sqrt{\lambda^2 - 1}$  and our assumption  $\lambda - 1 \gtrsim n^{-1/9} \log n$ . Combining the above bound with (118) gives

$$|\gamma_t \pi_t - \mu_t \alpha_t| \lesssim \alpha_t^2 \wedge \frac{1}{\alpha_t^2} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \lesssim \left( \frac{\log^4 n}{n(\lambda-1)^3} \right)^{1/4}, \quad (121)$$

where the last inequality follows from the elementary fact that  $\min\{a, b\} \leq \sqrt{ab}$ .

- Taking inequalities (115), (118) and (119) collectively with the fact that  $\mu_t \alpha_t^2 \asymp \alpha_t$  (cf. (113)) yields

$$|K_1(x) - K_2(x)| \lesssim \left| \frac{\pi_t}{\sqrt{n}} - \alpha_t \right| |\alpha_t + x| \lesssim \frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \cdot (|x| + \alpha_t), \quad (122)$$

and

$$\begin{aligned} |J_1(x) - J_2(x)| &\leq |\gamma_t \pi_t - \mu_t \alpha_t| |x| + 2 \left| \frac{1}{\sqrt{n}} \gamma_t \pi_t^2 - \mu_t \alpha_t^2 \right| \left| \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) \right| \\ &\quad + \mu_t \alpha_t^2 \left| \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh(\alpha_t(\alpha_t + x)) \right| \\ &\lesssim \left( \frac{\log^4 n}{n(\lambda-1)^3} \right)^{1/4} |x| + \frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} (|x| + \alpha_t) \\ &\lesssim \left( \frac{\log^4 n}{n(\lambda-1)^3} \right)^{1/4} |x| + \frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}. \end{aligned} \quad (123)$$

Equipped the above relations, we are positioned to control the right-hand side of expression (112). Combining (122) and (123) directly yields

$$\int |I_1(x)|^2 \varphi(dx) = \int [J_2(x) K_2(x)]^2 \varphi(dx) + O\left(\frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right) + O\left(\left(\frac{\log^4 n}{n(\lambda-1)^3}\right)^{1/4} + \frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}\right)$$

$$\begin{aligned}
&= \int \left[ \mu_t \alpha_t x + 2\mu_t \alpha_t^2 \tanh(\alpha_t(\alpha_t + x)) \right]^2 \left[ 1 - \tanh^2(\alpha_t(\alpha_t + x)) \right]^2 \varphi(dx) \\
&\quad + O \left( \left( \frac{\log^4 n}{n(\lambda-1)^3} \right)^{1/4} + \frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \right), \tag{124}
\end{aligned}$$

where the first line invokes the simple observations that  $|J_i(x)| \lesssim |x| + \alpha_t$  and  $|K_i(x)| \lesssim 1$  for  $i = 1, 2$ . Similarly, one can derive in the same manner that

$$\int |I_2(x)|^2 \varphi(dx) = \mu_t^2 \alpha_t^2 \int \left[ 1 - \tanh^2(\alpha_t(\alpha_t + x)) \right]^2 \varphi(dx) + O \left( \left( \frac{\log^4 n}{n(\lambda-1)^3} \right)^{1/4} + \frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \right). \tag{125}$$

As it turns out, the main terms in the above two identities satisfy

$$\int \left[ \mu_t \alpha_t x + 2\mu_t \alpha_t^2 \tanh(\alpha_t(\alpha_t + x)) \right]^2 \left[ 1 - \tanh^2(\alpha_t(\alpha_t + x)) \right]^2 \varphi(dx) \leq 1 - c\alpha_t^2, \tag{126a}$$

$$\mu_t^2 \alpha_t^2 \int \left[ 1 - \tanh^2(\alpha_t(\alpha_t + x)) \right]^2 \varphi(dx) \leq 1 - c\alpha_t^2, \tag{126b}$$

which we shall justify momentarily. Combine these results with (104) to conclude that: if  $\sqrt{\lambda^2 - 1} n^{-0.1} \lesssim \alpha_t \lesssim \sqrt{\lambda^2 - 1}$ , then

$$\kappa_t^2 \leq 1 - c\alpha_t^2 + O \left( \left( \frac{\log^4 n}{n(\lambda-1)^3} \right)^{1/4} + \frac{1}{\alpha_t} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \right) = 1 + o \left( \frac{\lambda-1}{\log n} \right), \tag{127}$$

where the last inequality follows by recognizing that

$$\left( \frac{\log^4 n}{n(\lambda-1)^3} \right)^{1/4} + \frac{1}{\sqrt{\lambda-1} n^{-0.1}} \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} = o \left( \frac{\lambda-1}{\log n} \right)$$

under our assumption  $\lambda - 1 \gtrsim n^{-1/9} \log n$ .

**Proof of relation (126).** To proceed, consider the problem of estimating  $v^*$  (which obeys  $v_i^* \sim \text{Unif}\{\pm \frac{1}{n}\}$ ) from the noisy observation  $Y = \alpha_t v^* + g$ , where  $g \sim \mathcal{N}(0, \frac{1}{n} I_n)$ . As alluded to previously, the Bayes-optimal estimate (or minimum mean square estimator (MMSE)) is given by

$$\mathbb{E}[v^* | Y] = \tanh(\sqrt{n} \alpha_t Y),$$

which satisfies (due to its optimality)

$$\text{Cor}(v^*, f(Y)) \leq \text{Cor}(v^*, \mathbb{E}[v^* | Y]), \tag{128}$$

for any measurable function  $f$ ; here  $\text{Cor}(\cdot, \cdot)$  denotes the correlation of two random vectors. In particular, the Bayes-optimal estimator outperforms the identity estimator (i.e.,  $f(Y) = Y$ ), so that (128) translates to

$$\begin{aligned}
\frac{\mathbb{E}[\langle v^*, \alpha_t v^* + g \rangle]}{\sqrt{\mathbb{E}[\| \alpha_t v^* + g \|_2^2]}} &\leq \frac{\mathbb{E}[\langle v^*, \tanh(\sqrt{n} \alpha_t (\alpha_t v^* + g)) \rangle]}{\sqrt{\mathbb{E}[\| \tanh(\sqrt{n} \alpha_t (\alpha_t v^* + g)) \|_2^2]}} = \frac{\int \tanh(\alpha_t(\alpha_t + x)) \varphi(dx)}{\sqrt{\int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx)}} \\
&= \sqrt{\int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx)}, \tag{129}
\end{aligned}$$

where the fist equality holds due to the symmetry of  $\varphi(\cdot)$ , and the second equality holds since  $\int \tanh(\alpha^2 + \alpha x)\varphi(dx) = \int \tanh^2(\alpha^2 + \alpha x)\varphi(dx)$  (see Deshpande et al. (2017, Appendix B.2)). As a consequence, the above relation implies that

$$\frac{\alpha_t}{\sqrt{\alpha_t^2 + 1}} = \frac{\mathbb{E}[\langle v^*, \alpha_t v^* + g \rangle]}{\sqrt{\mathbb{E}[\| \alpha_t v^* + g \|_2^2]}} \leq \sqrt{\int \tanh^2(\alpha_t(\alpha_t + x))\varphi(dx)} = \frac{1}{\mu_t}, \quad (130)$$

which in turns reveals that

$$\mu_t \alpha_t \leq \sqrt{\alpha_t^2 + 1} =: \gamma.$$

Armed with this relation, we can conclude that

$$\begin{aligned} & \max \left\{ \int [\mu_t \alpha_t x + 2\mu_t \alpha_t^2 \tanh(\alpha_t(\alpha_t + x))] [1 - \tanh^2(\alpha_t(\alpha_t + x))] \varphi(dx), \right. \\ & \quad \left. \mu_t^2 \alpha_t^2 \int [1 - \tanh^2(\alpha_t(\alpha_t + x))]^2 \varphi(dx) \right\} \\ & \leq \gamma^2 \max \left\{ \int [x + 2\alpha_t \tanh(\alpha_t(\alpha_t + x))] [1 - \tanh^2(\alpha_t(\alpha_t + x))] \varphi(dx), \right. \\ & \quad \left. \int [1 - \tanh^2(\alpha_t(\alpha_t + x))]^2 \varphi(dx) \right\} \\ & =: \kappa^2(\gamma, \alpha_t^2). \end{aligned}$$

As it turns out, this function  $\kappa^2(\cdot, \cdot)$  has been studied in Li and Wei (2022); more specifically, Li and Wei (2022, relation (272)) together with  $\gamma := \sqrt{\alpha_t^2 + 1}$  indicates that  $\kappa(\gamma, \alpha_t^2) \leq 1 - \frac{\gamma-1}{12}$ , and hence

$$\kappa^2(\gamma, \alpha_t^2) \leq 1 - \frac{\gamma-1}{12} \leq 1 - c\alpha_t^2 \quad (131)$$

for some constant  $c > 0$ . Here, notice that we view  $\gamma$  and  $\alpha_t^2$  as  $\lambda$  and  $\tau$  respectively in Li and Wei (2022, relation (272)). Putting everything together completes the proof of the required relation (126).

**Case III:**  $\alpha_t \gtrsim \sqrt{\lambda^2 - 1}$  and  $t < \frac{cn(\lambda-1)^5}{\log^2 n}$ . The calculation of  $\kappa_t$  in this case follows from similar arguments as in Li and Wei (2022, Section D.3.4.). The only difference lies in the computing the parameters  $\pi_t$  and  $\gamma_t$ , which was done in Li and Wei (2022, Lemma 14) therein but requires a different proof here. Specifically, we aim to show that

$$\pi_t = (1 + o(\lambda - 1))\alpha_t \sqrt{n} \quad \text{and} \quad \gamma_t^{-2} = (1 + o(\lambda - 1))n \int \tanh(\alpha_t(\alpha_t + x))\varphi(dx). \quad (132)$$

If these two relations were valid, then one could follow the argument in (Li and Wei, 2022, Section D.3.4.) verbatim to demonstrate that

$$\kappa_t \leq 1 - \frac{1}{15}(\lambda - 1),$$

as claimed.

We now present how to prove relation (132). In view of the equation (20a) of Lemma 2, we have

$$\begin{aligned} \pi_t &= \alpha_t \sqrt{n} \left( 1 + \frac{1}{\alpha_t^2} O\left( \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \right) \right) = \alpha_t \sqrt{n} \left( 1 + O\left( \sqrt{\frac{(t + \log^3 n / (\lambda - 1)) \log n}{n(\lambda - 1)^2}} \right) \right) \\ &= \alpha_t \sqrt{n} \left( 1 + O\left( \sqrt{\frac{(\lambda - 1)^2}{\log n}} \right) \right) = \alpha_t \sqrt{n} (1 + o(\lambda - 1)) \end{aligned}$$

under the condition  $t \lesssim \frac{n(\lambda-1)^5}{\log^2 n}$  and the assumption (25). In addition, with the same analysis as inequality (116), we can guarantee that

$$\begin{aligned} \int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx) &= \int \tanh^2 (\alpha_t (\alpha_t + x)) \varphi(dx) + O \left( \sqrt{\frac{(t + \log^3 n / (\lambda - 1)) \log n}{n \alpha_t^2 (\lambda - 1)^2}} \right) \\ &= \int \tanh (\alpha_t (\alpha_t + x)) \varphi(dx) + O \left( \sqrt{\frac{(\lambda - 1)^2}{\log n}} \right) = \int \tanh (\alpha_t (\alpha_t + x)) \varphi(dx) + o(\lambda - 1). \end{aligned}$$

Then according to equation (20b), we can reach

$$\begin{aligned} \gamma_t^{-2} &= n \left( \int \tanh(\alpha_t (\alpha_t + x)) \varphi(dx) + o(\lambda - 1) \right) + n \alpha_t^2 (1 + o(\lambda - 1)) O \left( \sqrt{\frac{(t + \log^3 n / (\lambda - 1)) \log n}{n}} \right) \\ &= (1 + o(\lambda - 1)) n \int \tanh(\alpha_t (\alpha_t + x)) \varphi(dx), \end{aligned}$$

where the last equality follows from the fact that  $\int \tanh(\alpha_t (\alpha_t + x)) \varphi(dx) \asymp \alpha_t^2 \asymp 1$  (see relation (113)). This establishes the claim (132).

#### A.4 Proof of Claim (43)

This subsection aims to establish the advertised decomposition (43). To do so, recall that  $\{\eta_i(x_i)\}_{1 \leq i \leq t-1}$  spans the same linear space as  $\{z_i\}_{1 \leq i \leq t-1}$  (see (37) and (38b)). It is important to notice that  $\{\eta_1(x_1), \dots, \eta_{t-1}(x_{t-1})\}$  are almost orthogonal to each other, thus forming a set of near-orthonormal basis; this property is summarized in the lemma below, whose proof is provided in Section A.8.

**Lemma 5.** *Suppose that the assumptions of Theorem 1 hold. With probability at least  $1 - O(n^{-11})$ , we have*

$$\left\| \sum_{i=1}^t w_i \eta_i(x_i) \right\|_2 = (1 + o(1)) \|w\|_2 \quad (133)$$

simultaneously for all  $t \leq \tau_0$  and all  $w = [w_i]_{1 \leq i \leq t} \in \mathbb{R}^t$ , where  $\tau_0$  is defined in (42).

In view of Lemma 5 and the fact that  $\xi_t \in \text{span}(U_{t-1}) = \text{span}\{\eta_1(x_1), \dots, \eta_{t-1}(x_{t-1})\}$  (cf. (37)), one can write  $\xi_t$  as a linear combination of  $\{\eta_i(x_i)\}_{1 \leq i \leq t-1}$  as follows:

$$\xi_t = \sum_{k=1}^{t-1} \gamma_t^k \eta_k(x_k), \quad \text{with } \gamma_t = [\gamma_t^k]_{1 \leq k < t} \in \mathbb{R}^{t-1} \text{ obeying } \|\gamma_t\|_2 \asymp \|\xi_t\|_2. \quad (134)$$

Armed with this decomposition, we intend to prove that

$$\begin{aligned} \alpha_{t+1} &= \lambda v^{\star \top} \eta_t(x_t) = \lambda v^{\star \top} \eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_t^k \eta_k(x_k) \right) \\ &= \lambda v^{\star \top} \eta_t(v_t) + O \left( \frac{\log^{2.5} n}{n^{3/4} (\lambda - 1)^{1.5}} \right), \end{aligned} \quad (135)$$

which shall be done as follows.

- In order to see this, first note that  $\eta_t(\cdot)$  is a Lipschitz function with Lipschitz constant  $O(1)$  (see Lemma 3). Therefore, for every  $t \lesssim \frac{\log n}{\lambda - 1}$  we have

$$\left| v^{\star \top} \eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_t^k \eta_k(x_k) \right) - v^{\star \top} \eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_t^k \eta_k(v_k) \right) \right|$$

$$\leq \left\| \eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(x_k) \right) - \eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right) \right\|_2 \lesssim \sum_{k=1}^{t-1} |\gamma_{t-1}^k| \|\eta_k(x_k) - \eta_k(v_k)\|_2.$$

In view of the decomposition (134) and the Cauchy-Schwarz inequality, we can further obtain

$$\begin{aligned} & \left| v^{\star \top} \eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(x_k) \right) - v^{\star \top} \eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right) \right| \lesssim \sum_{k=1}^{t-1} |\gamma_{t-1}^k| \|\eta_k(x_k) - \eta_k(v_k)\|_2 \\ & \lesssim \sum_{k=1}^{t-1} |\gamma_{t-1}^k| \|\xi_{k-1}\|_2 \leq \|\gamma_{t-1}\|_2 \left( \sum_{k=1}^{t-1} \|\xi_{k-1}\|_2^2 \right)^{1/2} \\ & \asymp \|\xi_{t-1}\|_2 \left( \sum_{k=1}^{t-1} \|\xi_{k-1}\|_2^2 \right)^{1/2} \lesssim \sqrt{\frac{t^3 \log n}{n}} \cdot \left( \sum_{k=1}^{t-1} \frac{k^3 \log n}{n} \right)^{1/2} \lesssim \frac{\log^{4.5} n}{n(\lambda-1)^{3.5}}, \end{aligned}$$

where the last line invokes  $\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}$  (cf. (36)) and  $t \lesssim \frac{\log n}{\lambda-1}$ .

- In addition, when  $|\alpha_t| \lesssim \sqrt{\lambda-1} n^{-1/4}$ , we know that  $\left\| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_\infty \lesssim \sqrt{\frac{t \log n}{n}}$  conditioned on the event  $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$  (defined in Lemma 1 with  $\delta = O(n^{-10})$ ). It therefore guarantees that

$$|v_{t,i}| = \left| \alpha_t v_i^\star + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_{k,i} \right| \leq \frac{|\alpha_t|}{\sqrt{n}} + \left| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_{k,i} \right| \lesssim \sqrt{\frac{t \log n}{n}}, \quad (136a)$$

$$\begin{aligned} \left| v_{t,i} + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_{k,i}) \right| & \lesssim |v_{t,i}| + \sum_{k=1}^{t-1} |\gamma_{t-1}^k| |v_{k,i}| \lesssim |v_{t,i}| + \|\gamma_{t-1}\|_2 \|\tilde{v}_{t-1,i}\|_2 \\ & \lesssim \sqrt{\frac{t \log n}{n}} + \sqrt{\frac{t^3 \log n}{n}} \cdot \sqrt{\frac{t^2 \log n}{n}} \lesssim \sqrt{\frac{t \log n}{n}}, \end{aligned} \quad (136b)$$

for every  $1 \leq i < t$ , where we denote  $\tilde{v}_{t-1,i} := (v_{1,i}, v_{2,i}, \dots, v_{t-1,i})$ . To see why (136b) is valid, we note that the first inequality applies Lemma 3, the second inequality results from the Cauchy-Schwarz inequality, whereas the last line makes use of (136a) and the fact  $\|\gamma_{t-1}\|_2 \asymp \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}$  (cf. (134) and (36)). In addition, given that  $t \lesssim \log n / (\lambda-1)$  and  $\lambda-1 \gtrsim n^{-1/9} \log n$ , we have  $t^8 \lesssim n / \log n$ . Repeating the argument for inequality (166) in the proof of Lemma 5, one can ensure that for any  $2 \leq k \leq 14$ ,

$$\sum_{i=1}^n |v_{t,i}|^k = \sum_{i=1}^t |v_{t,(i)}|^k + \sum_{i=t+1}^n |v_{t,(i)}|^k \lesssim \left( \frac{\log n}{n} \right)^{k/2-1}, \quad (137a)$$

and

$$\sum_{i=1}^n \left| v_{t,i} + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_{k,i}) \right|^k \lesssim \left( 1 + t^{2k} \left( \frac{\log n}{n} \right)^{k/2} \right) \left( \frac{\log n}{n} \right)^{k/2-1} \lesssim \left( \frac{\log n}{n} \right)^{k/2-1}. \quad (137b)$$

Here, we have made the observation that

$$\begin{aligned} & \sum_{i=1}^n \left| \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_{k,i}) \right|^k \\ & \lesssim \sum_{i=1}^n \|\gamma_{t-1}\|_2^k \left( \sum_{k=1}^{t-1} \eta_k^2(v_{k,i}) \right)^{k/2} \lesssim \sum_{i=1}^n \left( \frac{t^3 \log n}{n} \right)^{k/2} \left( \sum_{k=1}^{t-1} v_{k,i}^2 \right)^{k/2} \\ & \lesssim \left( \frac{t^3 \log n}{n} \right)^{k/2} \sum_{i=1}^n \left( \sum_{k=1}^{t-1} v_{k,(i)}^2 \right)^{k/2} \end{aligned}$$

$$\begin{aligned}
&\lesssim \left( \frac{t^3 \log n}{n} \right)^{k/2} \cdot t \left( \frac{t^2 \log n}{n} \right)^{k/2} + \left( \frac{t^3 \log n}{n} \right)^{k/2} \left( \sum_{k=1}^{t-1} v_{k,(t+1)}^2 \right)^{k/2-1} \cdot \sum_{i=t+1}^n \left( \sum_{k=1}^{t-1} v_{k,(i)}^2 \right) \\
&\lesssim t^{\frac{5k}{2}+1} \left( \frac{\log n}{n} \right)^k + t^{2k} \left( \frac{\log n}{n} \right)^{k-1} \\
&\lesssim t^{2k} \left( \frac{\log n}{n} \right)^{k-1},
\end{aligned}$$

where the second line uses the fact  $\|\gamma_{t-1}\|_2 \asymp \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}$  (cf. (134) and (36)) and Lemma 3, the ante-penultimate line invokes inequality (136); the penultimate line follows from the fact that  $\|v_k\|_2 \lesssim 1$  (see e.g. (87)) and conditional on event  $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$ ,

$$|v_{k,(t+1)}| \leq \frac{|\alpha_k|}{\sqrt{n}} + \left| \sum_{i=1}^{k-1} \beta_{k-1}^i \phi_i \right|_{(t+1)} \lesssim \sqrt{\frac{\log n}{n}};$$

and the last line follows from the fact that  $t^{k/2+1} \lesssim t^8 \lesssim n/\log n$ . Therefore, combining Lemma 3 with expression (136) gives

$$\begin{aligned}
&\eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right) - \eta_t(v_t) \\
&= (1 - c_0) \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) + O(\pi_t^2) \cdot \left[ \left( v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right)^3 - (v_t)^3 \right] + c_x
\end{aligned} \tag{138}$$

for some vectors  $c_x \in \mathbb{R}^n$ , where the parameters obey  $\pi_t \lesssim (\lambda - 1)^{-3/4} n^{1/4} \log n$  and  $c_0 \lesssim \frac{\log^4 n}{\sqrt{n}(\lambda - 1)^3}$ .

Here, the last equation makes use of the fact that

$$\begin{aligned}
\|c_x\|_2 &\lesssim \left\| \frac{n|v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k)|^5 \log^4 n}{(\lambda - 1)^3} \right\|_2 + \left\| \frac{n|v_t|^5 \log^4 n}{(\lambda - 1)^3} \right\|_2 \\
&\lesssim \frac{n \log^4 n}{(\lambda - 1)^3} \sqrt{\left( \frac{\log n}{n} \right)^4} \lesssim \frac{\log^6 n}{n(\lambda - 1)^3},
\end{aligned} \tag{139}$$

where the property (137) is invoked with  $k = 10$ . Next, observe that

$$\begin{aligned}
&\left\| \left( v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right)^3 - (v_t)^3 \right\|_2 \\
&= \left\| \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \circ \left( (v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k))^2 + (v_t)^2 + (v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k)) \circ v_t \right) \right\|_2 \\
&\lesssim \left\| \max_{1 \leq k \leq t-1} \eta_k(v_k) \circ \left( (v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k))^2 + (v_t)^2 \right) \right\|_2 \cdot \sqrt{t} \|\xi_{t-1}\|_2 \\
&\lesssim \max_{1 \leq k \leq t-1} \|\eta_k(v_k)\|_\infty \cdot \left( \left\| (v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k))^2 \right\|_2 + \left\| (v_t)^2 \right\|_2 \right) \cdot \sqrt{t} \|\xi_{t-1}\|_2 \\
&\lesssim \frac{t \log n}{n} \cdot \|\xi_{t-1}\|_2 \lesssim \frac{\log^4 n}{n^{1.5}(\lambda - 1)^{2.5}},
\end{aligned}$$

where the last line can be obtained by invoking property (136) and (137) with  $k = 4$ . Here, we have used the facts that  $\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}$  (cf. (36)),  $t \lesssim \frac{\log n}{\lambda - 1}$  and

$$\left\| v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right\|_2 \leq \|v_t\|_2 + \left\| \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right\|_2 \leq 1 + \sqrt{t} \|\gamma_{t-1}\|_2 \leq 1 + \sqrt{\frac{t^4 \log n}{n}} \lesssim 1.$$

Putting these together, we arrive at

$$\left\| \eta_t \left( v_t + \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right) - \eta_t(v_t) - (1 - c_0) \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k) \right\|_2 = O \left( \frac{\log^6 n}{n(\lambda - 1)^3} \right).$$

- Finally, it is sufficient for us to consider  $v^{\star\top} \sum_{k=1}^{t-1} \gamma_{t-1}^k \eta_k(v_k)$  which shall be controlled as follows:

$$\begin{aligned} \left| \sum_{k=1}^{t-1} \gamma_{t-1}^k v^{\star\top} \eta_k(v_k) \right| &\leq \left| \sum_{k=1}^{t-1} \gamma_{t-1}^k [v^{\star\top} \eta_k(x_k) + O(\|\xi_{k-1}\|_2)] \right| \\ &= \left| \sum_{k=1}^{t-1} \gamma_{t-1}^k \left( \frac{\alpha_{k+1}}{\lambda} + O\left(\sqrt{\frac{k^3 \log n}{n}}\right) \right) \right| \\ &\lesssim \sqrt{t} \|\gamma_{t-1}\|_2 \cdot \frac{\sqrt{\lambda - 1}}{n^{1/4}} \asymp \sqrt{t} \|\xi_{t-1}\|_2 \cdot \frac{\sqrt{\lambda - 1}}{n^{1/4}} \lesssim \frac{\log^{2.5} n}{n^{3/4}(\lambda - 1)^{1.5}}. \end{aligned}$$

Putting the above three inequalities together yields the desired bound (135).

Built upon expression (135), we now proceed to the proof of claim (43). To begin with, let us recall that  $v_t := \alpha_t v^{\star} + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k$ . If we define  $g_{t-1} := v^{\star\top} \phi_{t-1} \sim \mathcal{N}(0, \frac{1}{n})$  and  $\tilde{\beta}_{t-1} := [\beta_{t-1}^1, \dots, \beta_{t-1}^{t-2}]$ , some direct algebra thus leads to

$$\begin{aligned} |v^{\star\top} (v_t - \alpha_t v^{\star} - \phi_{t-1})| &= \left| v^{\star\top} \left[ \sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k - (1 - \beta_{t-1}^{t-1}) \phi_{t-1} \right] \right| \\ &\leq \left| \sum_{k=1}^{t-2} \beta_{t-1}^k g_k \right| + |1 - \beta_{t-1}^{t-1}| |g_{t-1}| \\ &= \left| \sum_{k=1}^{t-2} \beta_{t-1}^k g_k \right| + \frac{1 - (\beta_{t-1}^{t-1})^2}{1 + |\beta_{t-1}^{t-1}|} |g_{t-1}| \\ &\lesssim \|\tilde{\beta}_{t-1}\|_2 \sqrt{\sum_{k=1}^{t-2} (g_k)^2} + \|\tilde{\beta}_{t-1}\|_2^2 |g_{t-1}| \lesssim \|\tilde{\beta}_{t-1}\|_2 \sqrt{\frac{t \log n}{n}} \\ &\lesssim \sqrt{\frac{t \log n}{n}} \cdot \left( \frac{t \sqrt{\lambda - 1}}{n^{1/4}} + \frac{t \log^4 n}{\sqrt{n(\lambda - 1)^3}} \right) \lesssim \frac{\log^2 n}{n^{3/4}(\lambda - 1)} + \frac{\log^6 n}{n(\lambda - 1)^3}, \end{aligned}$$

where the last inequality comes from the bound (162) in the proof of Lemma 5 and the condition  $t \lesssim \frac{\log n}{\lambda - 1}$ . By virtue of the above calculations, we can deduce that

$$v^{\star\top} v_t = \alpha_t + g_{t-1} + O \left( \frac{\log^2 n}{n^{3/4}(\lambda - 1)} + \frac{\log^6 n}{n(\lambda - 1)^3} \right).$$

In fact, a direct application of Lemma 3 further leads to the following claim:

$$\begin{aligned} |v^{\star\top} (\eta_t(v_t) - v_t)| &\lesssim \left| v^{\star\top} \left[ c_0 v_t + \pi_t^2(v_t \circ v_t \circ v_t) + O\left(\frac{n \log^4 n}{(\lambda - 1)^3} |v_t|^5\right) \right] \right| \\ &\lesssim \frac{\log^4 n}{n^{3/4}(\lambda - 1)}, \end{aligned} \tag{140}$$

whose the proof of the last inequality is postponed to the end of this subsection.

To summarize, taking the above results collectively and using the relation (135), we arrive at

$$\alpha_{t+1} = \lambda v^{\star\top} v_t + \lambda v^{\star\top} (\eta_t(v_t) - v_t) + O \left( \frac{\log^{2.5} n}{n^{3/4}(\lambda - 1)^{1.5}} \right)$$

$$= \lambda \alpha_t + \lambda g_{t-1} + O\left(\frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}}\right). \quad (141)$$

Therefore, invoking the above relation recursively leads to our desired decomposition:

$$\alpha_{t+1} = \lambda^{t-k+1} \alpha_k + \sum_{i=1}^{t-k+1} \lambda^i g_{t-i} + O\left(\sum_{i=1}^{t-k+1} \lambda^i \frac{\log^4 n}{n^{3/4}(\lambda-1)^{1.5}}\right)$$

for any  $1 \leq k \leq t$ .

**Proof of inequality (140).** In order to establish inequality (140), let us first make note of the following simple properties: with probability at least  $1 - O(n^{-11})$ ,

$$\begin{aligned} v^{*\top}(v^* \circ v^* \circ v^*) &= \frac{1}{n}; \\ v^{*\top}\left(v^* \circ v^* \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right) &= \frac{1}{n} v^{*\top}\left(\sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right) \lesssim \frac{1}{n}; \\ v^{*\top}\left(v^* \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right) &= \frac{1}{n} \left\| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_2^2 \asymp \frac{1}{n}; \\ v^{*\top}\left(\sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right) &\lesssim \frac{t \log n}{n} v^{*\top}\left(\sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right) \lesssim \sqrt{\frac{t^3 \log^3 n}{n^3}}. \end{aligned}$$

We remind the readers that  $v_i^* \sim \text{Unif}(\pm \frac{1}{\sqrt{n}})$  and we have invoked Lemma 1.

Next, recall  $v_t := \alpha_t v^* + \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k$  to obtain

$$\begin{aligned} |v^{*\top}(v_t \circ v_t \circ v_t)| &= \left| \alpha_t^3 v^{*\top}(v^* \circ v^* \circ v^*) + 3\alpha_t^2 v^{*\top}\left(v^* \circ v^* \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right) \right. \\ &\quad \left. + 3\alpha_t v^{*\top}\left(v^* \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right) + v^{*\top}\left(\sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \circ \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k\right) \right| \\ &\lesssim \frac{\alpha_t^3}{n} + \frac{\alpha_t^2}{n} + \frac{\alpha_t}{n} + \sqrt{\frac{t^3 \log^3 n}{n^3}} \\ &\lesssim \frac{1}{n^{5/4}}, \end{aligned}$$

where the last line holds as long as  $\alpha_t \lesssim \sqrt{\lambda-1} n^{-1/4}$ . Consequently, in order to derive (140), it suffices to notice (137),  $c_0 \lesssim \frac{\log^4 n}{\sqrt{n}(\lambda-1)^3}$ ,  $\pi_t \lesssim \frac{n^{1/4}}{(\lambda-1)^{3/4}}$  and

$$|v^{*\top} v_t| = \left| \alpha_t + v^{*\top}\left(\sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k\right) \right| \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}} + \sqrt{\frac{t \log n}{n}} \lesssim \frac{\sqrt{\lambda-1}}{n^{1/4}}.$$

## A.5 Proof of Claim (53)

For notational simplicity, we assume without loss of generality that  $\alpha_t > 0$  throughout this proof. Before delving into the proof of claim (53), let us recall Lemma 2 to obtain

$$\frac{\pi_t}{\alpha_t \sqrt{n}} = 1 + O\left(\frac{1}{\alpha_t^2} \left( \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \right) \wedge \frac{1}{\alpha_t} \left( \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \right)^{1/2}\right)$$

$$= 1 + O\left(\frac{\log^2 n}{\alpha_t^2 \sqrt{(\lambda-1)^3 n}} \wedge \left(\frac{\log^2 n}{\alpha_t^2 \sqrt{(\lambda-1)^3 n}}\right)^{1/2}\right),$$

where we have used  $\|\xi_t\|_2 \leq \sqrt{\frac{t^3 \log n}{n}}$  (see (36)) and  $t \lesssim \frac{\log n}{\lambda-1}$ . In turn, this implies

$$\left| \frac{\pi_t^2}{\alpha_t^2 n} - 1 \right| \alpha_t^2 \lesssim \frac{\log^2 n}{\sqrt{(\lambda-1)^3 n}}. \quad (142)$$

Now, let us move on to establish a recursive relation of  $\alpha_t$ . Recalling the definition (3) of  $\eta_t$  and Theorem 2, one sees that

$$\begin{aligned} \alpha_{t+1} &= \lambda v^{\star \top} \int \eta_t \left( \alpha_t v^{\star} + \frac{1}{\sqrt{n}} x \right) \varphi_n(dx) + \Delta_{\alpha,t} \\ &= \lambda \gamma_t v^{\star \top} \int \tanh \left( \pi_t \left( \alpha_t v^{\star} + \frac{1}{\sqrt{n}} x \right) \right) \varphi_n(dx) + \Delta_{\alpha,t} \\ &= \lambda \gamma_t \sqrt{n} \int \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx) + \Delta_{\alpha,t}, \end{aligned} \quad (143)$$

where the last equality holds by symmetry of  $\varphi(\cdot)$ , namely,

$$\frac{1}{\sqrt{n}} \int \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx) = - \frac{1}{\sqrt{n}} \int \tanh \left( \frac{\pi_t}{\sqrt{n}} (-\alpha_t + x) \right) \varphi(dx).$$

We note that similar analysis as for relation (113) leads to  $\int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx) \asymp \frac{\pi_t^2}{n}$ . Combining this result with Lemma 2 and (36), we arrive at

$$\gamma_t^{-2} = n \left( 1 + O\left(\sqrt{\frac{t^3 \log n}{n}}\right) \right) \int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx). \quad (144)$$

Taking (143) and (144) together, we arrive at

$$\alpha_{t+1} = \left( 1 + O\left(\sqrt{\frac{t^3 \log n}{n}}\right) \right) \frac{\lambda \int \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx)}{\left[ \int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx) \right]^{1/2}} + \Delta_{\alpha,t}. \quad (145)$$

To prove claim (53), it then suffices to control  $\int \tanh \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx)$  and  $\int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx)$ . Towards this goal, we find it helpful to first make several observations. Define two functions:

$$\begin{aligned} f(z) &:= \frac{1}{z} \tanh(zy) - \tanh(y), \\ g(z) &:= \frac{1}{z^2} \tanh^2(zy) - \tanh^2(y). \end{aligned}$$

The Taylor expansion of  $\tanh(zy)$  gives

$$\begin{aligned} f'(z) &= -\frac{1}{z^2} [\tanh(zy) - zy + zy \tanh^2(zy)] = -\frac{2}{3} zy^3 + O(z^3 y^5), \\ g'(z) &= -\frac{2 \tanh(zy)}{z^3} [\tanh(zy) - zy + zy \tanh^2(zy)] = \frac{1}{3} zy^4 + O(z^3 y^6), \end{aligned}$$

which leads the following relation by direct calculation

$$f(z) = \int_1^z f'(t) dt = -\frac{1}{3} (z^2 - 1) y^3 + O((z^4 - 1) y^5),$$

$$g(z) = \int_1^z g'(t)dt = \frac{1}{6}(z^2 - 1)y^4 + O((z^4 - 1)y^6).$$

By taking  $z = \frac{\pi_t}{\alpha_t \sqrt{n}}$ ,  $y = \alpha_t(\alpha_t + x)$ , we can see that

$$\begin{aligned} \frac{\alpha_t \sqrt{n}}{\pi_t} \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh(\alpha_t(\alpha_t + x)) &= -\frac{1}{3}\left(\frac{\pi_t^2}{\alpha_t^2 n} - 1\right)\alpha_t^3(\alpha_t + x)^3 + O\left(\frac{\pi_t^4}{\alpha_t^4 n^2} - 1\right)\alpha_t^5(\alpha_t + x)^5, \\ \frac{\alpha_t^2 n}{\pi_t^2} \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh^2(\alpha_t(\alpha_t + x)) &= \frac{1}{6}\left(\frac{\pi_t^2}{\alpha_t^2 n} - 1\right)\alpha_t^4(\alpha_t + x)^4 + O\left(\frac{\pi_t^4}{\alpha_t^4 n^2} - 1\right)\alpha_t^6(\alpha_t + x)^6. \end{aligned}$$

Hence, we can conclude that

$$\begin{aligned} &\left| \int \frac{\alpha_t \sqrt{n}}{\pi_t} \tanh\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh(\alpha_t(\alpha_t + x)) \varphi(dx) \right| \\ &= \left| \int \frac{1}{3}\left(\frac{\pi_t^2}{\alpha_t^2 n} - 1\right)\alpha_t^3(\alpha_t + x)^3 \varphi(dx) \right| + \int O\left(\frac{\pi_t^4}{\alpha_t^4 n^2} - 1\right)\alpha_t^5(\alpha_t + x)^5 \varphi(dx) \\ &\lesssim \left| \frac{\pi_t^2}{\alpha_t^2 n} - 1 \right| \alpha_t^4. \end{aligned}$$

Similarly, we can show that

$$\left| \int \frac{\alpha_t^2 n}{\pi_t^2} \tanh^2\left(\frac{\pi_t}{\sqrt{n}}(\alpha_t + x)\right) - \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx) \right| \lesssim \left| \frac{\pi_t^2}{\alpha_t^2 n} - 1 \right| \alpha_t^4. \quad (146)$$

Substituting these relations into (145), we arrive at

$$\begin{aligned} \alpha_{t+1} &= \lambda \left( 1 + O\left(\sqrt{\frac{t^3 \log n}{n}}\right) \right) \frac{\int \tanh(\alpha_t(\alpha_t + x)) \varphi(dx) + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^4\right)}{\left[\int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx) + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^4\right)\right]^{1/2}} + \Delta_{\alpha,t} \\ &= \lambda \left( 1 + O\left(\sqrt{\frac{t^3 \log n}{n}}\right) \right) \frac{\int \tanh(\alpha_t(\alpha_t + x)) \varphi(dx)}{\left[\int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx)\right]^{1/2}} \cdot \frac{1 + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^2\right)}{1 + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^2\right)} + \Delta_{\alpha,t} \\ &= \lambda \left[ \int \tanh^2(\alpha_t^2 + \alpha_t x) \varphi(dx) \right]^{-1/2} \int \tanh(\alpha_t^2 + \alpha_t x) \varphi(dx) + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^3 + \sqrt{\frac{t^3 \log n}{n}} \alpha_t + \Delta_{\alpha,t}\right), \end{aligned} \quad (147)$$

where we make use of the fact that (see, (113))

$$\int \tanh(\alpha_t^2 + \alpha_t x) \varphi(dx) = \int \tanh^2(\alpha_t^2 + \alpha_t x) \varphi(dx) \asymp \alpha_t^2.$$

In addition, (130) ensures that

$$\frac{\alpha_t}{\sqrt{\alpha_t^2 + 1}} \leq \sqrt{\int \tanh^2(\alpha_t(\alpha_t + x)) \varphi(dx)} = \left[ \int \tanh^2(\alpha_t^2 + \alpha_t x) \varphi_n(dx) \right]^{-1/2} \int \tanh(\alpha_t^2 + \alpha_t x) \varphi(dx),$$

which in turn gives

$$\alpha_{t+1} \geq \frac{\lambda \alpha_t}{\sqrt{\alpha_t^2 + 1}} + O\left(\left|\frac{\pi_t^2}{\alpha_t^2 n} - 1\right| \alpha_t^3 + \sqrt{\frac{t^3 \log n}{n}} \alpha_t + |\Delta_{\alpha,t}|\right).$$

To finish up, putting the above results together with (142) leads to

$$\alpha_{t+1} \geq \frac{\lambda \alpha_t}{\sqrt{\alpha_t^2 + 1}} + o((\lambda - 1)\alpha_t) + O(\Delta_{\alpha,t}) \quad (148)$$

where we again invoke the assumption that  $\lambda - 1 \gtrsim n^{-1/9} \log n$ . This concludes the proof of claim (53).

## A.6 Proof of Claim (56)

Consider the regime where

$$|\alpha_t| < (\lambda - 1)^{-3/4} n^{-1/4} \lesssim \sqrt{\lambda - 1} n^{-0.1}.$$

First, invoke property (22b) in Lemma 3 to ensure that

$$|v^* \eta_t(x_t) - v^* \eta_t(v_t)| \lesssim |v^* \xi_{t-1}| + (\lambda - 1)n^{-0.2}(\log n) \|\xi_{t-1}\|_2. \quad (149)$$

To further bound (149), note that  $\xi_{t-1}$  admits the following decomposition in terms of  $\{\eta_k(x_k)\}$ :

$$\xi_{t-1} = \sum_{k=1}^{t-2} \gamma_{t-1}^k \eta_k(x_k), \quad \text{with } \gamma_{t-1} = [\gamma_{t-1}^k]_{1 \leq k \leq t-2} \in \mathbb{R}^{t-2} \text{ obeying } \|\gamma_{t-1}\|_2 \lesssim \|\xi_{t-1}\|_2;$$

the proof of this claim can be found in Section A.4 (see Lemma 5 therein and its proof). In view of this relation, we can apply (149) and the Cauchy-Schwarz inequality to reach

$$\begin{aligned} |v^* \eta_t(x_t) - v^* \eta_t(v_t)| &\lesssim \left| \sum_{k=1}^{t-2} \gamma_{t-1}^k v^* \eta_k(x_k) \right| + (\lambda - 1)n^{-0.2}(\log n) \|\xi_{t-1}\|_2 \\ &\lesssim \|\gamma_{t-1}\|_2 \left( \sum_{k=1}^{t-2} (v^* \eta_k(x_k))^2 \right)^{1/2} + (\lambda - 1)n^{-0.2}(\log n) \|\xi_{t-1}\|_2 \\ &\lesssim \sqrt{t} \|\xi_{t-1}\|_2 \max_{\tau_0 \leq s \leq t} |\alpha_s| + (\lambda - 1)n^{-0.2}(\log n) \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t \log n}{n}}, \end{aligned}$$

provided that  $t \lesssim \frac{\log n}{\lambda - 1}$  and  $\max_{\tau_0 \leq s \leq t} |\alpha_s| \lesssim \sqrt{\lambda - 1} n^{-0.1}$ . Here the last inequality invokes  $\|\xi_t\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}$  (see (36)). Substitution into (52) yields

$$|\Delta_{\alpha,t}| \lesssim \sqrt{\frac{t \log n}{n}} \ll (\lambda - 1) |\alpha_t|, \quad (150)$$

given  $|\alpha_t| \gtrsim \sqrt{\lambda - 1} n^{-1/4}$  and  $\lambda - 1 \gtrsim n^{-1/9} \log n$ . It thus completes the proof of the relation (56).

## A.7 Proof of Claim (59)

Throughout this section, we assume without loss of generality that  $\alpha_t > 0$ . As computed in Section A.5 for relation (53), applying Lemma 2 reveals that

$$\begin{aligned} \left| \frac{\pi_t^2}{\alpha_t^2 n} - 1 \right| \alpha_t^2 &\lesssim \left( \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \right) \wedge \alpha_t \left( \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \right)^{1/2} \lesssim \alpha_t \left( \frac{(\lambda - 1)^3}{\log n} \right)^{1/4}, \\ \gamma_t^{-2} &= n \left( 1 + O \left( \sqrt{\frac{(\lambda - 1)^3}{\log n}} \right) \right) \int \tanh^2 \left( \frac{\pi_t}{\sqrt{n}} (\alpha_t + x) \right) \varphi(dx), \\ |\Delta_{\alpha,t}| &\lesssim \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}} \lesssim \sqrt{\frac{(\lambda - 1)^3}{\log n}}, \end{aligned}$$

given the inductive assumptions  $\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{(\lambda - 1)^3}{\log n}}$  when  $t \lesssim \frac{n(\lambda - 1)^5}{\log^2 n}$  and  $\alpha_t \gtrsim \sqrt{\lambda^2 - 1}$ . Now in view of similar calculations for (147), we can deduce that

$$\begin{aligned} \alpha_{t+1} &= \lambda \left[ \int \tanh(\alpha_t^2 + \alpha_t x) \varphi_n(dx) \right]^{1/2} + O \left( \left| \frac{\pi_t^2}{\alpha_t^2 n} - 1 \right| \alpha_t^3 + \frac{\lambda - 1}{\sqrt{\log n}} \alpha_t + \Delta_{\alpha,t} \right) \\ &\geq \frac{\lambda \alpha_t}{\sqrt{1 + \alpha_t^2}} + o \left( (\lambda - 1)^{3/4} \alpha_t^2 + (\lambda - 1) \alpha_t \right) \end{aligned} \quad (151)$$

where we have made use of the fact that  $\sqrt{\frac{(\lambda-1)^3}{\log n}} \ll (\lambda-1)\alpha_t$ .

We then demonstrate that this relation (151) together with a little algebra indicates that  $\alpha_{t+1} \geq \frac{1}{2}\sqrt{\lambda^2 - 1}$ . Specifically, consider the following two cases separately.

- First, consider the case where  $\alpha_t \leq \frac{2}{3}\sqrt{\lambda^2 - 1}$ . Akin to inequality (54), relation (151) implies the existence of some constant  $c > 0$  such that

$$\alpha_{t+1} \geq (1 + c(\lambda-1))\alpha_t + o((\lambda-1)\alpha_t) \geq \alpha_t \geq \frac{1}{2}\sqrt{\lambda^2 - 1}.$$

- Otherwise, consider the case where  $\alpha_t > \frac{2}{3}\sqrt{\lambda^2 - 1}$ . Recognizing the fact that  $\frac{\lambda\alpha_t}{\sqrt{1+\alpha_t^2}}$  is monotonically increasing in  $\alpha_t$ , we arrive at

$$\alpha_{t+1} \geq (1 + c(\lambda-1))\frac{2}{3}\sqrt{\lambda^2 - 1} + o((\lambda-1)^{3/4}) \geq \frac{1}{2}\sqrt{\lambda^2 - 1}.$$

Thus, this completes the proof of our desired bound (25).

## A.8 Proof of Lemma 5

Throughout the proof, we work with the event that  $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$  (defined in Lemma 1 with  $\delta = O(n^{-11})$ ), which holds true with probability at least  $1 - O(n^{-11})$ . On this event, one has  $\|\phi_t\|_\infty \lesssim \sqrt{\frac{\log n}{n}}$  and

$$\|x_t\|_\infty \leq \|\alpha_t v^\star\|_\infty + \left\| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_\infty + \|\xi_{t-1}\|_2 \lesssim \frac{|\alpha_t|}{\sqrt{n}} + \sqrt{\frac{t \log n}{n}} + \sqrt{\frac{t^3 \log n}{n}} \lesssim \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}} \quad (152)$$

for any  $t \leq \tau_0$ , where we remind the readers that (see (36), the property  $\|\beta_{t-1}\|_2 = 1$ , the definition of  $\mathcal{E}_1$ , and the definition (41) of  $\tau_0$ )

$$|\alpha_t| \lesssim \sqrt{\lambda-1} n^{-1/4}, \quad \left\| \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k \right\|_\infty \lesssim \sqrt{\frac{t \log n}{n}} \quad \text{and} \quad \|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}} \quad (153)$$

as long as  $t \lesssim \frac{\log n}{\lambda-1}$  (see (42)).

To show that  $\{\eta_1(x_1), \dots, \eta_t(x_t)\}$  forms a near-orthogonal basis, one strategy is to show that  $\eta_i(x_i) \approx \phi_{i-1}$  for each  $1 \leq i \leq t \leq \tau_0$ , which in turn implies that

$$\left\| \sum_{i=1}^t w_i \eta_i(x_i) \right\|_2 \approx \left\| \sum_{i=1}^t w_i \phi_{i-1} \right\|_2 \asymp \|w\|_2, \quad (154)$$

given that  $\phi_k \sim \mathcal{N}(0, \frac{1}{n}I_n)$  are independent Gaussian vectors; here, we introduce  $\phi_0 := x_1 \sim \mathcal{N}(0, \frac{1}{n}I_n)$  for notational convenience. Guided by this intuition, we first use the triangle inequality to derive that

$$\begin{aligned} & \left| \left\| \sum_{i=1}^t w_i \eta_i(x_i) \right\|_2 - \left\| \sum_{i=1}^t w_i \beta_{i-1}^{i-1} \phi_{i-1} \right\|_2 \right| \\ & \leq \left\| \sum_{i=1}^t w_i [\eta_i(x_i) - \eta_i(\beta_{i-1}^{i-1} \phi_{i-1})] \right\|_2 + \left\| \sum_{i=1}^t w_i [\eta_i(\beta_{i-1}^{i-1} \phi_{i-1}) - \beta_{i-1}^{i-1} \phi_{i-1}] \right\|_2 \end{aligned} \quad (155)$$

for any vector  $w \in \mathbb{R}^t$ . In order to bound these terms, we proceed with the following three steps.

- Let us first consider the difference between  $\eta_t(x_t)$  and  $x_t$ . Invoking property (22c) in Lemma 3 allows us to express

$$\eta_t(x_t) = (1 - c_0) \left( x_t - \frac{\pi_t^2}{3} x_t \circ x_t \circ x_t + c_{x_t} \right),$$

where  $c_0 \lesssim \frac{\log^2 n}{\sqrt{n(\lambda-1)^3}}$ ,  $\pi_t \lesssim \frac{n^{1/4} \log n}{(\lambda-1)^{3/4}}$ , and  $c_{x_t}$  is a vector obeying (cf. (152))

$$\|c_{x_t}\|_\infty \lesssim \frac{n\|x_t\|_\infty^5 \log^4 n}{(\lambda-1)^3} \lesssim \frac{\log^{14} n}{n^{3/2}(\lambda-1)^{10.5}}. \quad (156)$$

Then one has

$$\|\eta_t(x_t) - x_t\|_2 \lesssim c_0\|x_t\|_2 + \frac{1}{3}\pi_t 2\|x_t \circ x_t \circ x_t\|_2 + \sqrt{n}\|c_{x_t}\|_\infty. \quad (157)$$

We then claim that

$$\|\eta_t(x_t) - x_t\|_2 \lesssim \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}; \quad (158)$$

to streamline the presentation, the proof of this claim is deferred to the end of this section. Applying the same argument once gain also leads to

$$\|\eta_t(\beta_{t-1}^{t-1}\phi_{t-1}) - \beta_{t-1}^{t-1}\phi_{t-1}\|_2 \lesssim \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}. \quad (159)$$

- Combining relations (158) and (159) and invoking the triangle inequality, we obtain

$$\begin{aligned} \|\eta_t(x_t) - \eta_t(\beta_{t-1}^{t-1}\phi_{t-1})\|_2 &\leq \|\eta_t(x_t) - x_t\|_2 + \|x_t - \beta_{t-1}^{t-1}\phi_{t-1}\|_2 + \|\eta_t(\beta_{t-1}^{t-1}\phi_{t-1}) - \beta_{t-1}^{t-1}\phi_{t-1}\|_2 \\ &= \|x_t - \beta_{t-1}^{t-1}\phi_{t-1}\|_2 + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \\ &= \left\| \alpha_t v^\star + \sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k + \xi_t \right\|_2 + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \\ &\leq \|\alpha_t v^\star\|_2 + \left\| \sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k \right\|_2 + \|\xi_t\|_2 + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \\ &\leq \left\| \sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k \right\|_2 + |\alpha_t| + O\left(\sqrt{\frac{t^3 \log n}{n}}\right) + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right), \end{aligned} \quad (160)$$

where the last line makes use of (153). Let us denote  $\tilde{\beta}_{t-1} := (\beta_{t-1}^1, \dots, \beta_{t-1}^{t-2}) \in \mathbb{R}^{t-2}$ , obtained by removing the last entry of  $\beta_{t-1}$ . According to Lemma 1, we note that with probability  $1 - O(n^{-11})$ ,

$$\left\| \sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k \right\|_2 - \|\tilde{\beta}_{t-1}\|_2 \leq \|\tilde{\beta}_{t-1}\|_2 \cdot \sup_{a=[a_k]_{1 \leq k < t-2} \in \mathcal{S}^{t-3}} \left\| \left\| \sum_{k=1}^{t-2} a_k \phi_k \right\|_2 - 1 \right\| \lesssim \sqrt{\frac{t \log n}{n}} \|\tilde{\beta}_{t-1}\|_2,$$

which in turn implies that

$$\left(1 - O\left(\sqrt{\frac{t \log n}{n}}\right)\right) \|\tilde{\beta}_{t-1}\|_2 \leq \left\| \sum_{k=1}^{t-2} \beta_{t-1}^k \phi_k \right\|_2 \leq \left(1 + O\left(\sqrt{\frac{t \log n}{n}}\right)\right) \|\tilde{\beta}_{t-1}\|_2.$$

As a consequence, we can further control the right-hand side of (160) by

$$\|\eta_t(x_t) - \eta_t(\beta_{t-1}^{t-1}\phi_{t-1})\|_2 \leq \left(1 + O\left(\sqrt{\frac{t \log n}{n}}\right)\right) \|\tilde{\beta}_{t-1}\|_2 + O\left(|\alpha_t| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right). \quad (161)$$

- Our next step is concerned with bounding the term  $\|\tilde{\beta}_{t-1}\|_2$ . First, recall that  $\beta_t$  corresponds to the linear coefficients of  $\eta_t(x_t)$  when projected to the linear space  $U_t$  (see (39) and (38b)). We can thus write

$$\|\tilde{\beta}_t\|_2 = \|U_{t-1}^\top \eta_t(x_t)\|_2$$

$$\begin{aligned}
&\leq \|U_{t-1}^\top [\eta_t(x_t) - \eta_t(\beta_{t-1}^{t-1}\phi_{t-1})]\|_2 + \|U_{t-1}^\top [\eta_t(\beta_{t-1}^{t-1}\phi_{t-1}) - \beta_{t-1}^{t-1}\phi_{t-1}]\|_2 + \|U_{t-1}^\top (\beta_{t-1}^{t-1}\phi_{t-1})\|_2 \\
&\leq \left(1 + O\left(\sqrt{\frac{t \log n}{n}}\right)\right) \|\tilde{\beta}_{t-1}\|_2 + O\left(|\alpha_t| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) + \|U_{t-1}^\top (\beta_{t-1}^{t-1}\phi_{t-1})\|_2 \\
&= \left(1 + O\left(\sqrt{\frac{t \log n}{n}}\right)\right) \|\tilde{\beta}_{t-1}\|_2 + O\left(|\alpha_t| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) + O\left(\sqrt{\frac{t \log n}{n}}\right) \\
&= \left(1 + O\left(\sqrt{\frac{t \log n}{n}}\right)\right) \|\tilde{\beta}_{t-1}\|_2 + O\left(|\alpha_t| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right).
\end{aligned}$$

Here, the third line follows from (159) and (161); the penultimate line holds due to the independence between  $\phi_{t-1}$  and  $U_{t-1}$  (see the properties below display (40)) and hence  $U_{t-1}^\top \phi_{t-1} \sim \mathcal{N}(0, \frac{1}{n} I_t)$ ; and the last line holds as long as  $t \lesssim \frac{\log n}{\lambda-1}$ . Recognizing that  $\|\tilde{\beta}_1\|_2 = \sqrt{1 - \|\beta_1\|_2^2} = 0$ , we can apply the above relation recursively to yield

$$\begin{aligned}
\|\tilde{\beta}_t\|_2 &\leq \sum_{\tau=1}^t \left(1 + O\left(\sqrt{\frac{t \log n}{n}}\right)\right)^{t-1-\tau} \cdot O\left(|\alpha_\tau| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \\
&\lesssim \sum_{\tau=1}^t \left(|\alpha_\tau| + \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \lesssim \frac{t\sqrt{\lambda-1}}{n^{1/4}} + \frac{t \log^3 n}{\sqrt{n(\lambda-1)^3}},
\end{aligned}$$

where the penultimate inequality follows from the fact  $\sqrt{\frac{t \log n}{n}} \lesssim \frac{1}{t}$  as  $t \lesssim \frac{\log n}{\lambda-1}$ , and the last inequality uses  $|\alpha_t| \lesssim \sqrt{\lambda-1} n^{-1/4}$  for  $t \leq \tau_0$  (see the definition of  $\tau_0$  in (41)). Under the assumption that  $\lambda-1 \gtrsim n^{-1/9} \log n$ , we can further obtain

$$\|\tilde{\beta}_t\|_2 \lesssim \frac{t\sqrt{\lambda-1}}{n^{1/4}}. \quad (162)$$

Plugging this relation into (161) and using the condition  $|\alpha_t| \lesssim \sqrt{\lambda-1} n^{-1/4}$  ( $\forall t \leq \tau_0$ ) give

$$\|\eta_t(x_t) - \eta_t(\beta_{t-1}^{t-1}\phi_{t-1})\|_2 \lesssim \frac{t\sqrt{\lambda-1}}{n^{1/4}}. \quad (163)$$

It is worth noting that the inequality (162) also implies

$$|\beta_t^t| = \sqrt{\|\beta_t\|_2^2 - \|\tilde{\beta}_t\|_2^2} = \sqrt{1 - \|\tilde{\beta}_t\|_2^2} = 1 - O\left(\frac{t^2(\lambda-1)}{n^{1/2}}\right).$$

To finish up, putting the above bounds together with expression (155), we conclude that

$$\begin{aligned}
&\left| \left\| \sum_{i=1}^t w_i \eta_i(x_i) \right\|_2 - \left\| \sum_{i=1}^t w_i \phi_{i-1} \right\|_2 \right| \\
&\leq (1 - \beta_{t-1}^{t-1}) \left\| \sum_{i=1}^t w_i \phi_{i-1} \right\|_2 + \left\| \sum_{i=1}^t w_i [\eta_i(x_i) - \eta_i(\beta_{i-1}^{i-1}\phi_{i-1})] \right\|_2 + \left\| \sum_{i=1}^t w_i [\eta_i(\beta_{i-1}^{i-1}\phi_{i-1}) - \beta_{i-1}^{i-1}\phi_{i-1}] \right\|_2 \\
&= O\left(\frac{t^2(\lambda-1)}{n^{1/2}}\right) \left\| \sum_{i=1}^t w_i \phi_{i-1} \right\|_2 + O\left(\frac{t\sqrt{\lambda-1}}{n^{1/4}}\right) \cdot \sqrt{t} \|w\|_2 + O\left(\frac{\log^3 n}{\sqrt{n(\lambda-1)^3}}\right) \cdot \sqrt{t} \|w\|_2 \\
&= O\left(\frac{t^{3/2}\sqrt{\lambda-1}}{n^{1/4}} + \sqrt{\frac{t \log^6 n}{n(\lambda-1)^3}}\right) \|w\|_2 = o(\|w\|_2),
\end{aligned}$$

where the last relation results from the facts  $t < \tau_0 \lesssim \log n / (\lambda-1)$  and the assumption  $\lambda-1 \geq n^{-1/9}$ . Finally, observing  $\|\sum_{i=1}^t w_i \phi_{i-1}\|_2 = (1 + O(\sqrt{\frac{t \log n}{n}})) \|w\|_2$ , we reach

$$\left\| \sum_{i=1}^t w_i \eta_i(x_i) \right\|_2 = (1 + o(1)) \|w\|_2,$$

which completes the proof of our desired bound.

**Proof of inequality (158).** For any fixed integer  $k \geq 2$  that does not scale with  $n$ , we can write

$$\sum_{i=1}^n |x_{t,i}|^k \lesssim \sum_{i=1}^n |\alpha_t v_i^*|^k + \sum_{i=1}^n |u_{t,i}|^k + \sum_{i=1}^n |\xi_{t-1,i}|^k, \quad \text{with } u_t := \sum_{k=1}^{t-1} \beta_{t-1}^k \phi_k.$$

Let us bound each term separately. Firstly, recalling that  $\|\xi_{t-1}\|_2 \lesssim \sqrt{\frac{t^3 \log n}{n}}$  (cf. (36)) gives

$$\sum_{i=1}^n |\xi_{t-1,i}|^k \leq \|\xi_{t-1}\|_2^k \lesssim \left( \frac{t^3 \log n}{n} \right)^{k/2}. \quad (164)$$

Secondly, on the event  $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$  (see Lemma 1), we see that  $\|u_t\|_2 \lesssim 1$  and  $\|u_t\|_\infty \leq \sqrt{(t \log n)/n}$ . This in turn gives

$$\begin{aligned} \sum_{i=1}^n |u_{t,i}|^k &= \sum_{i \leq t} |u_{t,(i)}|^k + \sum_{i>t} |u_{t,(i)}|^k \stackrel{(*)}{\lesssim} t \left( \frac{t \log n}{n} \right)^{k/2} + \left( \frac{\log n}{n} \right)^{k/2-1} \sum_{i>t} |u_{t,(i)}|^2 \\ &\lesssim t \left( \frac{t \log n}{n} \right)^{k/2} + \left( \frac{\log n}{n} \right)^{k/2-1} \\ &\lesssim \left( 1 + \frac{t^{k/2+1} \log n}{n} \right) \left( \frac{\log n}{n} \right)^{k/2-1}, \end{aligned} \quad (165)$$

with  $x_{(i)}$  denoting the  $i$ -th largest entry of  $x$  (in magnitude). Here, to see why inequality  $(*)$  holds, we recall that on the event  $\{\phi_k\}_{k=1}^{t-1} \in \mathcal{E}$  (see Lemma 1), one has

$$\sup_{a \in \mathcal{S}^{t-2}} \sum_{i=1}^t \left| \sum_{k=1}^{t-1} a_k \phi_k \right|_{(i)}^2 \lesssim \frac{t \log n}{n},$$

which also guarantees that  $|u_{t,(t+1)}| \lesssim \sqrt{\log n/n}$ . Putting the above pieces together, we obtain

$$\sum_{i=1}^n |x_{t,i}|^k \lesssim n \left( \frac{1}{n(\lambda-1)} \right)^{3k/4} + \left( 1 + \frac{t^{k/2+1} \log n}{n} \right) \left( \frac{\log n}{n} \right)^{k/2-1} + \left( \frac{t^3 \log n}{n} \right)^{k/2} \lesssim \left( \frac{\log n}{n} \right)^2, \quad (166)$$

where the last inequality is valid if we take  $k = 6$ ,  $t \lesssim \log n/(\lambda-1)$  and assume  $\lambda-1 \gtrsim n^{-1/9} \log n$ . It therefore leads to

$$\pi_t^2 \|x_t \circ x_t \circ x_t\|_2 = \pi_t^2 \left( \sum_{i=1}^n |x_{t,i}|^6 \right)^{1/2} \lesssim \frac{n^{1/2} \log^2 n}{(\lambda-1)^{3/2}} \cdot \frac{\log n}{n} = \frac{\log^3 n}{\sqrt{n(\lambda-1)^3}},$$

where we have used the bound on  $\pi_t$  in Lemma 3 (the 3rd case). This together with (156), (157) and the fact  $c_0 \lesssim \frac{\log^2 n}{\sqrt{n(\lambda-1)^3}}$  concludes the proof of inequality (158).

## A.9 Proof of inequality (70)

Before proceeding, let us make several observations about  $\frac{\tau_t}{h(\tau_t)}$ . As discussed around Li and Wei (2022, display (254)), the sequence  $\tau_t$  with  $\tau_{t+1} = \lambda^2 h(\tau_t)$  is monotonically increasing, which implies that  $\frac{\tau_t}{h(\tau_t)} \leq \lambda^2$ . In addition, the optimality of the Bayes estimator (cf. (130)) implies that  $\frac{\tau_t}{h(\tau_t)} \leq \tau_t + 1$ . Combining these two observations, we obtain

$$\frac{\tau_t}{h(\tau_t)} \leq (\tau_t + 1) \wedge \lambda^2.$$

In view of the inductive assumption, we have  $\alpha_t^2 = (1 + o(1))\tau_t$  for  $t \gtrsim \varsigma$ . Hence, for every  $\tau$  obeying  $\min\{\tau_t, \alpha_t^2\} \leq \tau \leq \max\{\tau_t, \alpha_t^2\}$ , it holds that  $\tau = (1 + o(1))\tau_t$  with  $\tau_t \gtrsim \lambda^2 - 1$ . Define  $\mathcal{T}_2$  as in display (263) of Li and Wei (2022) such that

$$\mathcal{T}_2(s, \tau) := s^2 h'(\tau) = s^2 \int \left(1 + \frac{x}{2\sqrt{\tau}}\right) (1 - \tanh^2(\tau + \sqrt{\tau}x)) \varphi(dx). \quad (167)$$

Armed with this notation, we can bound the target quantity as

$$\frac{\tau_t}{h(\tau_t)} h'(\tau) \leq \mathcal{T}_2(\sqrt{\tau_t + 1} \wedge \lambda, \tau).$$

Therefore, it suffices to upper bound the right-hand side of the above inequality by  $1 - c(\lambda - 1)$ .

Towards this end, direct calculations yield

$$\frac{\mathcal{T}_2(\sqrt{\tau_t + 1} \wedge \lambda, \tau) - \mathcal{T}_2(\sqrt{\tau + 1} \wedge \lambda, \tau)}{\mathcal{T}_2(\sqrt{\tau + 1} \wedge \lambda, \tau)} = (\sqrt{\tau + 1} \wedge \lambda)^2 - (\sqrt{\tau_t + 1} \wedge \lambda)^2 = o(\lambda - 1). \quad (168)$$

Moreover, it has been proved numerically (see Figure 1 in Li and Wei (2022)) that

$$\mathcal{T}_2(\lambda, \tau) \leq 1 - (\lambda - 1), \quad \text{for } \lambda \in (0, 1.2] \text{ and } \tau > \sqrt{\lambda^2 - 1}.$$

Recognizing that  $\tau = (1 + o(1))\tau_t \gtrsim \lambda^2 - 1$ , we can deduce from the relation above that

$$\mathcal{T}_2(\sqrt{\tau + 1} \wedge \lambda, \tau) \leq 1 - ((\sqrt{\tau + 1} \wedge \lambda) - 1) = 1 - c_1(\lambda - 1) \quad (169)$$

for some universal constant  $c_1 > 0$ . Finally, putting relations (168) and (169) together, we arrive at

$$\frac{\tau_t}{h(\tau_t)} h'(\tau) \leq \mathcal{T}_2(\sqrt{\tau_t + 1} \wedge \lambda, \tau) = (1 + o(\lambda - 1))\mathcal{T}_2(\sqrt{\tau + 1} \wedge \lambda, \tau) \leq 1 - c(\lambda - 1)$$

for some universal constant  $c > 0$ . We have thus finished the proof of relation (70).

## B Proof of expression (17) and Corollary 1

To begin with, by definition (16) of  $u_t$ , one has

$$\|u_t\|_2 = \left\| \frac{1}{\lambda \sqrt{n(\alpha_t^2 + 1)}} \tanh(\pi_t x_t) \right\|_2 \leq \frac{1}{\lambda},$$

where we have used the fact that  $|\tanh(\pi_t x_t)| < 1$ . Therefore the quantity of interest  $\|v^* v^{*\top} - u_t u_t^\top\|_{\mathbb{F}}^2$  is uniformly upper bounded by a constant  $1 + \frac{1}{\lambda^4}$ . In addition, we find it helpful to observe that

$$\begin{aligned} \|v^* v^{*\top} - u_t u_t^\top\|_{\mathbb{F}}^2 &= \|v^*\|_2^4 - 2(v^{*\top} u_t)^2 + \|u_t\|_2^4 \\ &= 1 - 2 \frac{1}{n\lambda^2(\alpha_t^2 + 1)} (v^{*\top} \tanh(\pi_t x_t))^2 + \frac{1}{n^2 \lambda^4 (\alpha_t^2 + 1)^2} \|\tanh(\pi_t x_t)\|_2^4 \\ &= 1 - \frac{1}{\lambda^4} \left( \frac{2\alpha_{t+1}^2}{(\alpha_t^2 + 1)n\gamma_t^2} + \frac{1}{(\alpha_t^2 + 1)^2 n^2 \gamma_t^4} \right). \end{aligned} \quad (170)$$

To validate expression (17), it is sufficient to notice that  $\gamma_t^{-2} - n\alpha_t^2(\alpha_t^2 + 1) = o(1)$  with probability at least  $1 - O(n^{-10})$  (according to Lemma 2), which in turn leads to

$$\lim_{t \rightarrow \infty} \lim_{n \rightarrow \infty} \mathbb{E}[\|v^* v^{*\top} - u_t u_t^\top\|_{\mathbb{F}}^2] = 1 - \frac{\alpha^{*\top} \alpha}{\lambda^4}.$$

To prove Corollary 1, we again invoke Lemma 2 to demonstrate that

$$\begin{aligned}
\frac{1}{n}\gamma_t^{-2} &= \frac{1}{n}\pi_t^2 \left( \alpha_t^2 + 1 + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right) \right) \\
&= \left( \alpha_t^2 + O\left(\left(\|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right)^{1/2}\right) \right) \left( \alpha_t^2 + 1 + O\left(\frac{\pi_t^2}{n} + \|\xi_{t-1}\|_2 + \sqrt{\frac{t \log n}{n}}\right) \right) \\
&= (\alpha_t^2 + O(\delta^{1/2}))(\alpha_t^2 + 1 + O(\delta)) \\
&= \alpha_t^2(\alpha_t^2 + 1) + O(\delta^{1/2})
\end{aligned} \tag{171}$$

with  $\delta := \sqrt{\frac{t \log n}{n(\lambda-1)^2}} + \sqrt{\frac{\log^4 n}{n(\lambda-1)^3}}$ , where we plug in the bound on  $\|\xi_t\|_2$  as in expression (8c). Substituting the expression (171) into (170) yields

$$\|v^* v^{*\top} - u_t u_t^\top\|_F^2 = 1 - \frac{\alpha_t^2}{\lambda^4} (2\alpha_{t+1}^2 - \alpha_t^2 + O(\delta^{1/2})). \tag{172}$$

After an order of  $\frac{\log n}{\lambda-1}$  iterations, property (14) ensures that  $\alpha_t^2 - \alpha^{*2} = O(\sqrt{\frac{\log^4 n}{n(\lambda-1)^6}})$ . Putting everything together, we arrive at

$$\|v^* v^{*\top} - u_t u_t^\top\|_F^2 = 1 - \frac{\alpha^{*4}}{\lambda^4} + O\left(\sqrt{\frac{\log^4 n}{n(\lambda-1)^6}}\right),$$

which holds true with probability at least  $1 - O(n^{-10})$ .

## References

- Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of statistics*, 48(3):1452.
- Baik, J., Arous, G. B., and Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697.
- Bandeira, A. S., Kunisky, D., and Wein, A. S. (2019). Computational hardness of certifying bounds on constrained pca problems. *arXiv preprint arXiv:1902.07324*.
- Bayati, M. and Montanari, A. (2011a). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.
- Bayati, M. and Montanari, A. (2011b). The LASSO risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.
- Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815.
- Bu, Z., Klusowski, J. M., Rush, C., and Su, W. J. (2020). Algorithmic analysis and statistical estimation of SLOPE via approximate message passing. *IEEE Transactions on Information Theory*, 67(1):506–537.
- Cademartori, C. and Rush, C. (2023). A non-asymptotic analysis of generalized approximate message passing algorithms with right rotationally invariant designs. *arXiv preprint arXiv:2302.00088*.
- Cai, T. T. and Zhang, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89.
- Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

- Capitaine, M., Donati-Martin, C., and Féral, D. (2009). The largest eigenvalues of finite rank deformation of large wigner matrices: convergence and nonuniversality of the fluctuations. *The Annals of Probability*, 37(1):1–47.
- Celentano, M., Fan, Z., and Mei, S. (2021). Local convexity of the TAP free energy and AMP convergence for Z2-synchronization. *arXiv preprint arXiv:2106.11428*.
- Celentano, M. and Montanari, A. (2022). Fundamental barriers to high-dimensional regression with convex penalties. *The Annals of Statistics*, 50(1):170–196.
- Celentano, M., Montanari, A., and Wei, Y. (2020). The Lasso with general Gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*.
- Chen, Y. and Candès, E. J. (2018). The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Communications on Pure and Applied Mathematics*, 71(8):1648–1714.
- Chen, Y., Chi, Y., Fan, J., and Ma, C. (2019). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37.
- Chen, Y., Chi, Y., Fan, J., and Ma, C. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.
- Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2020). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM journal on optimization*, 30(4):3098–3121.
- Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*.
- Cheng, C., Wei, Y., and Chen, Y. (2021). Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Transactions on Information Theory*, 67(11):7380–7419.
- Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.
- Deshpande, Y., Abbe, E., and Montanari, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Information and Inference: A Journal of the IMA*, 6(2):125–170.
- Deshpande, Y., Montanari, A., and Richard, E. (2014). Cone-constrained principal component analysis. *Advances in Neural Information Processing Systems*, 27.
- Donoho, D. and Montanari, A. (2016). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.
- Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.
- Donoho, D. L., Maleki, A., and Montanari, A. (2010). Message passing algorithms for compressed sensing: I. motivation and construction. In *2010 IEEE information theory workshop on information theory (ITW 2010, Cairo)*, pages 1–5. IEEE.
- El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175.
- Fan, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197–224.
- Fan, Z., Mei, S., and Montanari, A. (2021). TAP free energy, spin glasses and variational inference. *The Annals of Probability*, 49(1):1–45.

- Feng, O. Y., Venkataramanan, R., Rush, C., and Samworth, R. J. (2022). A unifying tutorial on approximate message passing. *Foundations and Trends® in Machine Learning*, 15(4):335–536.
- Féral, D. and Péché, S. (2007). The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in mathematical physics*, 272(1):185–228.
- Gao, C. and Zhang, A. Y. (2022). SDP achieves exact minimax optimality in phase synchronization. *IEEE Transactions on Information Theory*.
- Ge, R., Jin, C., and Zheng, Y. (2017). No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR.
- Javanmard, A., Montanari, A., and Ricci-Tersenghi, F. (2016). Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223.
- Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327.
- Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.
- Keshavan, R., Montanari, A., and Oh, S. (2009). Matrix completion from noisy entries. *Advances in neural information processing systems*, 22.
- Lelarge, M. and Miolane, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probability Theory and Related Fields*, 173(3):859–929.
- Lesieur, T., Krzakala, F., and Zdeborová, L. (2017). Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(7):073403.
- Li, G. and Wei, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. *arXiv preprint arXiv:2208.03313*.
- Li, Y. and Wei, Y. (2021). Minimum  $\ell_1$ -norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*.
- Ma, C., Wang, K., Chi, Y., and Chen, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632.
- Ma, J., Xu, J., and Maleki, A. (2018). Optimization-based AMP for phase retrieval: The impact of initialization and  $\ell_2$ -regularization. *arXiv preprint arXiv:1801.01170*.
- Massart, P. (2007). *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer.
- Miolane, L. and Montanari, A. (2021). The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335.
- Mondelli, M. and Venkataramanan, R. (2021). PCA initialization for approximate message passing in rotationally invariant models. *Advances in Neural Information Processing Systems*, 34:29616–29629.
- Montanari, A. and Venkataramanan, R. (2021). Estimation of low-rank matrices via approximate message passing. *The Annals of Statistics*, 49(1):321–345.
- Péché, S. (2006). The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173.
- Perry, A., Wein, A. S., Bandeira, A. S., and Moitra, A. (2018). Message-passing algorithms for synchronization problems over compact groups. *Communications on Pure and Applied Mathematics*, 71(11):2275–2322.

- Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE.
- Rush, C. and Venkataraman, R. (2018). Finite sample analysis of approximate message passing algorithms. *IEEE Transactions on Information Theory*, 64(11):7264–7286.
- Singer, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36.
- Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.
- Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1-2):487–558.
- Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized  $m$ -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628.
- Wei, Y., Wainwright, M. J., and Guntuboyina, A. (2019). The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *The Annals of Statistics*, 47(2):994–1024.
- Wu, Y. and Zhou, K. (2022). Lower bounds for the convergence of tensor power iteration on random overcomplete models. *arXiv preprint arXiv:2211.03827*.
- Yan, Y., Chen, Y., and Fan, J. (2021). Inference for heteroskedastic PCA with missing data. *arXiv preprint arXiv:2107.12365*.
- Zhang, Y., Mondelli, M., and Venkataraman, R. (2022). Precise asymptotics for spectral methods in mixed generalized linear models. *arXiv preprint arXiv:2211.11368*.
- Zhong, X., Wang, T., and Fan, Z. (2021). Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. *arXiv preprint arXiv:2110.02318*.
- Zhong, Y. and Boumal, N. (2018). Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016.